

UNIVERSIDADE NOVE DE JULHO - UNINOVE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO
CONHECIMENTO

RICARDO PINTO FERREIRA

INTELIGÊNCIA COMPUTACIONAL NA PREVISÃO DO ABSENTEÍSMO E
IDENTIFICAÇÃO DE TENDÊNCIAS ABSENTEÍSTAS

SÃO PAULO
2019

RICARDO PINTO FERREIRA

**INTELIGÊNCIA COMPUTACIONAL NA PREVISÃO DO ABSENTEÍSMO E
IDENTIFICAÇÃO DE TENDÊNCIAS ABSENTEÍSTAS**

Tese de doutorado apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento - PPGI da Universidade Nove de Julho, como requisito parcial para obtenção do título de doutor em Informática e Gestão do Conhecimento.

Orientador: Prof. Dr. Renato José Sassi

**SÃO PAULO
2019**

Ferreira, Ricardo Pinto.

Inteligência computacional na previsão do absentéismo e identificação de tendências absenteístas. / Ricardo Pinto Ferreira. 2019. 200 f.

Tese (Doutorado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2019.

Orientador (a): Prof. Dr. Renato José Sassi.

1. Absenteísmo. 2. Gestão de pessoas. 3. Redes neurais artificiais. 4. Teoria dos Rough Sets. 5. Inteligência computacional.

I. Sassi, Renato José. II. Título.

CDU 004

“Se o conhecimento pode criar problemas, não é através da
ignorância que podemos solucioná-los” Isaac Asimov

DEDICATÓRIA

Dedico este trabalho aos meus pais João Pinto Ferreira e Durvalina Mello Ferreira (*in memoriam*), meus filhos Aleister e Arthur e, de maneira muito especial à minha esposa Andréa Martiniano da Silva.

AGRADECIMENTOS

A Deus, o “Maravilhoso Conselheiro”, que nos ilumina, nos fortalece e nos consola nos momentos mais difíceis. Devemos a Ele tudo o que temos recebido e tudo o que somos.

Ao Orientador Prof. Dr. Renato José Sassi, por sua incansável dedicação e paciência em compartilhar seu conhecimento.

À Universidade Nove de Julho e à Universidade Corporativa dos Correios pela bolsa de estudos no Programa de Pós-Graduação em Informática e Gestão do Conhecimento. Aos docentes e funcionários da Universidade Nove de Julho, em especial, ao Prof. Dr. André Felipe Henriques Librantz.

Aos membros da banca, Prof. Dr. Marco Antonio Silveira, Prof. Dr. Fabio Silva Lopes, Prof. Dr. Fellipe Silva Martins, Prof. Dr. André Felipe Henriques Librantz pelo pronto atendimento ao convite.

A todos os alunos do Programa de Pós-Graduação em Informática e Gestão do Conhecimento - PPGI da Universidade Nove de Julho, em especial, aos orientados do Prof. Dr. Renato José Sassi.

Àqueles que contribuíram, direta e indiretamente, para que este trabalho fosse realizado e, de maneira muito especial, ao meu irmão Márcio Pinto Ferreira.

RESUMO

O absenteísmo é um fenômeno definido como o não comparecimento do empregado ao trabalho, de forma habitual e com frequência regular, e, por conseguinte, como o não cumprimento das obrigações trabalhistas, conforme o programado. Entender e tratar as causas do absenteísmo têm sido um desafio para muitos gestores, dada a dimensão desse fenômeno, que engloba causas de fundo psicológico, físico e ambiental. A previsão do absenteísmo e a identificação de tendências absenteístas são importantes para reduzir as perdas da empresa e, ao mesmo tempo, para melhorar a qualidade de vida do empregado. Para isso, faz-se necessária uma análise das bases de dados, que armazenam informações sobre os empregados de uma empresa, por vários anos, o que abre espaço para a aplicação de técnicas de inteligência computacional, como as redes neurais artificiais. Diante disso, este estudo teve como objetivo aplicar técnicas da inteligência computacional na previsão do absenteísmo e na identificação de tendências absenteístas. A base de dados utilizada possui 50 atributos, com 2.403 registros de licenças médicas de 39 empregados, coletados durante o período de janeiro de 2008 a dezembro de 2017. Os experimentos computacionais foram realizados em duas fases: a fase 1, denominada Previsão do Absenteísmo, foi dividida em duas etapas; na etapa 1, aplicou-se a rede neural artificial do tipo *Multilayer Perceptron* (MLP), e na etapa 2, aplicou-se a Teoria dos *Rough Sets* para redução de atributos com o uso de dois métodos, o Algoritmo Genético e o Algoritmo de Johnson. Em seguida, aplicou-se a *Multilayer Perceptron*. Na fase 2, denominada Identificação de Tendências Absenteístas (etapa 3), foi utilizada a rede neural artificial do tipo *Self-Organizing Map* (SOM). Mediante uma comparação entre os resultados obtidos nas etapas 1 e 2, constatou-se que a MLP apresentou o erro experimental ligeiramente melhor do que as MLPs aplicadas na base de dados reduzida com a Teoria dos *Rough Sets*. No entanto, houve uma considerável redução do tempo de processamento dos experimentos computacionais na etapa 2. Vale ressaltar que os resultados das duas etapas apontaram positivamente para a previsão do absenteísmo. Na fase 2, etapa 3, os resultados gerados também apontaram positivamente para a identificação de tendências absenteístas por meio da avaliação de agrupamentos. Concluiu-se, então, que as técnicas de inteligência computacional, aplicadas para a previsão do absenteísmo e identificação de tendências absenteístas, permitiram atingir o objetivo aqui proposto, e mostraram-se como importantes técnicas para solucionar problemas complexos de absenteísmo, que afligem tanto as organizações quanto os empregados.

Palavras-chave: Absenteísmo, Gestão de Pessoas, Redes Neurais Artificiais, Teoria dos *Rough Sets*, Inteligência Computacional.

ABSTRACT

Absenteeism is considered a phenomenon defined as the non-attendance of the employee to work in a habitual way, with regular frequency and therefore the noncompliance of the obligations, as scheduled. Understanding and treating the causes of absenteeism has been a challenge, given the dimension of the phenomenon that encompasses psychological, physical and environmental causes. The prediction of absenteeism and the identification of absenteeism tendencies are important to reduce losses for the company and at the same time improve the quality of life of the employee. To this end, it is necessary to extract knowledge from databases that store information about employees of the company for several years, which opens space for the application of computational intelligence techniques, such as artificial neural networks. Thus, the objective of this work was to apply computational intelligence techniques in the prediction of absenteeism and in the identification of absenteeism tendencies. The database used is composed of 50 attributes with 2,403 medical license records from 39 employees collected during the period from January 2008 to December 2017. The computational experiments were carried out in two phases: Phase 1, called prediction absenteeism was In Phase 1, the artificial neural network of the type Multilayer Perceptron (MLP) was applied in Step 2 and in Step 2 the Rough Sets Theory was applied to reduce attributes using two reduction methods, the Genetic Algorithm and the Johnson Algorithm, and then applied the Multilayer Perceptron. In Phase 2, called the Self-Organizing Map artificial neural network, called Step 3. The comparison between the results obtained in Steps 1 and 2 made it possible to verify that the MLP presented the slightly better experimental error of the that the MLPs applied in the database reduced with the Rough Sets Theory. However, there was a considerable reduction in the processing time of the computational experiments in Step 2. It is noteworthy that the results of the two steps pointed positively to the prediction of absenteeism. In Phase 2, Step 3, identification of absenteeism tendencies with the Self-Organizing Map, the results generated also pointed positively to identify absenteeism tendencies by means of clustering evaluation. It is concluded that the computational intelligence techniques applied for the prediction of absenteeism and the identification of absenteeism tendencies have managed to reach the proposed objective and are presented as important techniques for the understanding and possible solution of this complex problem that afflicts both organizations employees.

Keywords: Absenteeism, Management, Artificial Neural Networks, Rough Sets Theory, Computational Intelligence.

SUMÁRIO

1. INTRODUÇÃO	20
1.1 Justificativa e motivação.....	22
1.2 Problema de pesquisa.....	23
1.3 Objetivos.....	24
1.3.1 Objetivo geral.....	24
1.3.2 Objetivos específicos.....	24
1.4 Delimitação do tema de pesquisa.....	24
1.5 Organização do trabalho.....	25
2. FUNDAMENTAÇÃO TEÓRICA.....	26
2.1 Absenteísmo.....	26
2.1.1 Estudos relacionados ao Absenteísmo.....	27
2.2 Descoberta de Conhecimento em Bases de dados.....	32
2.2.1 Fases do KDD.....	33
2.2.2 Tarefas do KDD.....	36
2.3 Técnicas da Inteligência Computacional.....	39
2.3.1 Redes Neurais Artificiais.....	39
2.3.1.1 <i>Multilayer Perceptron</i>	43
2.3.1.2 <i>Self-Organizing Map</i>	46
2.3.1.3 O Algoritmo de Aprendizado da Rede SOM.....	47
2.3.1.4 O Processo Competitivo da rede SOM.....	48
2.3.1.5 O Processo Cooperativo da rede SOM.....	52
2.3.1.6 O Processo Adaptativo Sináptico da rede SOM.....	56
2.3.1.7 Medidas de Qualidade da rede SOM.....	58
2.3.1.8 Teoria dos Conjuntos Aproximados.....	59
2.3.1.8.1 Principais Conceitos.....	60
2.3.2 Revisão Bibliométrica.....	72
3. MATERIAIS E MÉTODOS.....	79
3.1 Caracterização metodológica.....	79
3.2 Caracterização da empresa.....	80
3.3 Base de dados, <i>softwares</i> e plataforma de experimentos.....	81
3.4 Condução dos experimentos computacionais.....	87

4. APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS.....	90
4.1 Experimentos computacionais.....	90
4.2 Fase 1 - Etapa 1 (RNA): Aplicação de uma RNA na previsão do absentismo (Treinamento).....	91
4.3 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos <i>Rough Sets</i> e MLP na previsão do absentismo (Experimento 1 - Treinamento).....	109
4.4 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos <i>Rough Sets</i> e MLP na previsão do absentismo (Experimento 2 - Treinamento).....	119
4.5 Síntese dos resultados dos experimentos computacionais Fase 1 (Treinamento).....	130
4.6 Fase 1 - Etapa 1 (RNA): Aplicação de uma MLP na previsão do absentismo (Teste).....	131
4.7 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos <i>Rough Sets</i> com método de redução Algoritmo genético e MLP na previsão do absentismo (Experimento 1 - Teste).....	140
4.8 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos <i>Rough Sets</i> com método de redução Algoritmo de Johnson e MLP na previsão do absentismo (Experimento 2 - Teste).....	148
4.9 Síntese dos resultados dos experimentos computacionais Fase 1 - Etapas 1 e 2 (Teste).....	157
4.10 Fase 2 - Identificação de tendências absenteeístas - Etapa 3 (SOM).....	159
4.11 Síntese dos resultados dos experimentos computacionais Fase 2 - Etapa 3.....	166
5. CONCLUSÃO.....	169
5.1 Contribuição dos experimentos computacionais.....	169
5.2 Contribuição para as empresas.....	171
5.3 Contribuição para os empregados.....	171
5.4 Contribuição teórica e limitação.....	171
5.5 Contribuição da revisão bibliométrica.....	172
5.6 Contribuição para estudos futuros.....	173
5.7 Memorial de Publicações.....	174
5.8 Bases de dados cedidas para <i>UCI Machine Learning Repository</i>	175
5.9 Registro de programa de computador oriundo da pesquisa.....	177

REFERÊNCIAS BIBLIOGRÁFICAS.....	178
Anexo A – Relação dos 100 artigos selecionados na revisão bibliométrica.....	194
Anexo B – Certificado de Registro de Programa de Computador.....	200

LISTA DE FIGURAS

Figura 1 - Um modelo multinível da dinâmica social do fenômeno do absenteísmo.....	31
Figura 2 - Fases do processo de KDD.....	33
Figura 3 - Esquema simplificado da mineração de dados.....	35
Figura 4 - Representação simplificada de um neurônio biológico e um neurônio artificial....	40
Figura 5 - Estrutura básica da MLP.....	43
Figura 6 - Função sigmóide.....	45
Figura 7 - Mapa unidimensional e bidimensional retangular e hexagonal.....	46
Figura 8 - Grade bidimensional de neurônios da rede SOM.....	48
Figura 9 - Modelo da rede SOM.....	49
Figura 10 - Modelo de vizinhança da rede SOM.....	50
Figura 11 - Processo de ajuste do vetor de pesos do neurônio vencedor.....	51
Figura 12 - Distribuição em espaço bidimensional dos vetores.....	52
Figura 13 - Função de vizinhança gaussiana apresentada sob diferentes vistas.....	53
Figura 14 - Vizinhança topológica retangular.....	54
Figura 15 - Espaço Aproximado $A = (U,R)$ e $X \subseteq U$	61
Figura 16 - Aproximação Inferior de X	65
Figura 17 - Aproximação Superior de X	66
Figura 18 - Todas as regiões de X em A	67
Figura 19 - Distribuição dos artigos por país de origem.....	73
Figura 20 - Distribuição dos artigos pesquisados por aspecto comum.....	73
Figura 21 - Classificação dos artigos do ponto de vista dos procedimentos metodológicos....	76
Figura 22 - Classificação dos artigos do ponto de vista da forma de abordagem do problema.....	76
Figura 23 - Fases e etapas dos experimentos computacionais.....	88
Figura 24 - Etapas 1 e 2 dos experimentos computacionais que compõem a Fase 1.....	89
Figura 25 - Etapa 3 dos experimentos computacionais que compõem a Fase 2.....	89
Figura 26 - Realização dos experimentos da Fase 1, Etapa 1 da fase de Treinamento.....	91
Figura 27 - Topologia da MLP utilizada no experimento.....	92
Figura 28 - Fase de treinamento da MLP para a previsão do absenteísmo mensal medido em horas.....	94

Figura 29 - Fase de treinamento da MLP para a previsão do absenteísmo bimestral medido em horas.....	96
Figura 30 - Fase de treinamento da MLP para a previsão do absenteísmo trimestral medido em horas.....	98
Figura 31 - Fase de treinamento da MLP para a previsão do absenteísmo quadrimestral medido em horas.....	100
Figura 32 - Fase de treinamento da MLP para a previsão do absenteísmo semestral medido em horas.....	102
Figura 33 - Fase de treinamento da MLP para a previsão do absenteísmo anual medido em horas.....	104
Figura 34 - Fase de treinamento da MLP para a previsão do absenteísmo bianual medido em horas.....	106
Figura 35 - Curva do erro quadrático médio na fase de treinamento.....	108
Figura 36 - Realização dos experimentos da Fase 1, Etapa 2, Experimento 1 – Treinamento.....	109
Figura 37 - Topologia da MLP utilizada no Experimento 1.....	111
Figura 38 - Fase de treinamento da MLP para a previsão do absenteísmo mensal medido em horas.....	112
Figura 39 - Fase de treinamento da MLP para a previsão do absenteísmo trimestral medido em horas.....	114
Figura 40 - Fase de treinamento da MLP para a previsão do absenteísmo quadrimestral medido em horas.....	116
Figura 41 - Curva do erro quadrático médio na fase de treinamento.....	118
Figura 42 - Realização dos experimentos da Etapa 2 (Experimento 2 - Treinamento).....	120
Figura 43 - Topologia da MLP utilizada na Fase 1, Etapa 2 (Experimento 2 - Treinamento)..	121
Figura 44 - Fase de treinamento da MLP para a previsão do absenteísmo mensal medido em horas.....	123
Figura 45 - Fase de treinamento da MLP para a previsão do absenteísmo trimestral medido em horas.....	125
Figura 46 - Fase de treinamento da MLP para a previsão do absenteísmo quadrimestral medido em horas.....	127
Figura 47 - Curva do erro quadrático médio na fase de treinamento.....	129
Figura 48 - Fase de teste da MLP para a previsão do absenteísmo mensal medido em horas...	133

Figura 49 - Fase de teste da MLP para a previsão do absenteísmo trimestral medido em horas.....	135
Figura 50 - Fase de teste da MLP para a previsão do absenteísmo quadrimestral medido em horas.....	137
Figura 51 - Curva do erro quadrático médio na fase de treinamento e teste.....	139
Figura 52 - Fase de teste da MLP para a previsão do absenteísmo mensal medido em horas...	141
Figura 53 - Fase de teste da MLP para a previsão do absenteísmo trimestral medido em horas.....	143
Figura 54 - Fase de teste da MLP para a previsão do absenteísmo quadrimestral medido em horas.....	145
Figura 55 - Curva do erro quadrático médio na fase de treinamento e teste.....	147
Figura 56 - Fase de teste da MLP para a previsão do absenteísmo mensal medido em horas...	149
Figura 57 - Fase de teste da MLP para a previsão do absenteísmo trimestral medido em horas.....	151
Figura 58 - Fase de teste da MLP para a previsão do absenteísmo quadrimestral medido em horas.....	154
Figura 59 - Curva do erro quadrático médio na fase de treinamento e teste.....	156
Figura 60 - Realização dos experimentos da Fase 2.....	159
Figura 61 – Os três agrupamentos gerados pela rede SOM.....	160

LISTA DE TABELAS

Tabela 1 - Exemplo de um Sistema de Informação (S).....	62
Tabela 2 - Sistema de Decisão (Sistema de Informação com o Atributo de Decisão: Tendência).....	63
Tabela 3 - Sistema de Decisão com os elementos Ação2 e Ação3 indiscerníveis, com relação aos atributos condicionais.....	64
Tabela 4 - Sistema de Informação S sem o atributo Preço Máximo.....	69
Tabela 5 - Sistema de Informação S sem o atributo Nova Mínima.....	70
Tabela 6 - Matriz de Discernibilidade.....	70
Tabela 7 - Distribuição dos artigos por continente.....	73
Tabela 8 - Distribuição dos artigos por idioma.....	74
Tabela 9 - Distribuição dos artigos por base de dados.....	74
Tabela 10 - Distribuição dos artigos por período.....	75
Tabela 11 - Período e quantidade de publicações que compõem o estudo.....	79
Tabela 12 - Distribuição por tipo de publicação.....	80
Tabela 13 - Classificação Internacional de Doenças (CID).....	84
Tabela 14 - Lista dos seis motivos (categorias) sem CID.....	85
Tabela 15 - Lista de atributos que compõem a base de dados.....	86
Tabela 16 - Atributos de entrada da MLP (43 atributos).....	91
Tabela 17 - Parâmetros utilizados na MLP (43 neurônios de entrada).....	92
Tabela 18 - Redutos gerados utilizando algoritmo genético e a base reduzida.....	109
Tabela 19 - Atributos de entrada da MLP (17 atributos).....	110
Tabela 20 - Parâmetros utilizados na MLP (17 neurônios de entrada).....	110
Tabela 21 - Atributos de entrada da MLP (16 atributos).....	120
Tabela 22 - Parâmetros utilizados na MLP (16 neurônios de entrada).....	121
Tabela 23 - Distribuição da frequência dos acertos por classes (mensal).....	134
Tabela 24 - Distribuição da frequência dos acertos por classes (trimestral).....	136
Tabela 25 - Distribuição da frequência dos acertos por classes (quadrimestral).....	138
Tabela 26 - Distribuição da frequência dos acertos por classes (mensal).....	142
Tabela 27 - Distribuição da frequência dos acertos por classes (trimestral).....	144
Tabela 28 - Distribuição da frequência dos acertos por classes (quadrimestral).....	146
Tabela 29 - Distribuição da frequência dos acertos por classes (mensal).....	150

Tabela 30 - Distribuição da frequência dos acertos por classes (trimestral).....	152
Tabela 31 - Distribuição da frequência dos acertos por classes (quadrimestral).....	155
Tabela 32 - Representatividade dos três agrupamentos gerados pela rede SOM.....	160
Tabela 33 - Tendência do absenteísmo anual no agrupamento A1.....	162
Tabela 34 - Tendência do absenteísmo anual no agrupamento A2.....	163
Tabela 35 - Tendência do absenteísmo anual no agrupamento A3.....	164
Tabela 36 - Tendência do absenteísmo bianual no agrupamento A1.....	165
Tabela 37 - Tendência do absenteísmo bianual no agrupamento A2.....	165
Tabela 38 - Tendência do absenteísmo bianual no agrupamento A3.....	166

LISTA DE QUADROS

Quadro 1 - Atributos utilizados que identificam fatores de risco ao absenteísmo em trabalhadores de uma empresa de petróleo da Região Norte-Nordeste do Brasil entre 2007 e 2009.....	28
Quadro 2 - Descrição dos 50 atributos que compõem a base de dados.....	82
Quadro 3 - <i>Softwares</i> utilizados no trabalho.....	87
Quadro 4 - Fases dos experimentos computacionais.....	90
Quadro 5 - Síntese dos resultados dos experimentos computacionais da Fase 1 (Treinamento).....	130
Quadro 6 - Características dos experimentos computacionais da Fase 1 – Etapas 1 e 2 (Teste).....	157
Quadro 7 - Síntese dos resultados dos experimentos computacionais da Fase 1 – Etapa 1 e 2 (Teste).....	158
Quadro 8 - Características dos atributos de cada agrupamento gerado pela rede SOM.....	161
Quadro 9 - Síntese dos resultados dos experimentos computacionais da Fase 2 – Etapa 3 no período anual.....	167
Quadro 10 - Síntese dos resultados dos experimentos computacionais da Fase 2 – Etapa 3 no período bianual.....	168

LISTA DE ABREVIATURAS E SIGLAS

AG - Algoritmo Genético
BMU - *Best Match Unit*
CID - Código Internacional de Doenças
CN - Computação Natural
CPU - *Central Processing Unit*
DM - *Data Mining*
EQ - Erro de Quantização
ET – Erro Topográfico
GB - *Gigabyte*
IC - Inteligência Computacional
IMC - Índice de massa corporal
INPI - Instituto Nacional da Propriedade Industrial
KDD - *Knowledge Discovery in Databases*
MBP - *Multiple Back-Propagation*
MLP - *MultiLayer Perceptron*
OMS - Organização Mundial de Saúde
RAM - *Random Access Memory*
RBF - *Radial Basis Function*
RNA - Rede Neural Artificial
RNAs - Redes Neurais Artificiais
RS - *Rough Sets*
SOM - *Self-Organizing Maps*
SVM - *Support Vector Machines*
TB – *Terabyte*
WHO - *World Health Organization*

1. INTRODUÇÃO

As empresas vivenciam constantes avanços tecnológicos, característicos do processo de evolução com consequências benéficas para a sociedade. No entanto, esses avanços aumentam a competitividade empresarial, o que pode provocar mudanças no comportamento biológico, psicológico e social dos empregados, acarretando em sobrecarga física e mental (McEWEN, 2007; BHUI et al., 2012). Essas mudanças e tensões desencadeiam reações adversas, interferindo diretamente na qualidade de vida no trabalho e que podem levar ao fenômeno chamado de absenteísmo (JODAS e HADADD, 2009; LOHAUS e HABERMANN, 2019).

O absenteísmo é considerado um fenômeno definido como o não comparecimento do empregado ao trabalho de forma habitual, com frequência regular e conseqüentemente o não cumprimento das obrigações, conforme o programado, o que difere da falta ao trabalho em situação pontual isolada em função de sua frequência e habitualidade (JOHNS, 2008; PRATER e SMITH, 2011; LOHAUS e HABERMANN, 2019). Assim, para identificar o motivo do baixo rendimento dos seus empregados, faz-se necessário um estudo de suas causas e consequências que podem estar relacionadas ao absenteísmo (INOUE et al., 2008; BRATSBERG, FEVANG e RØED, 2013).

Segundo Tolbert et al. (2014) e Hargrave et al. (2008) por conta do absenteísmo, os prejuízos causados às empresas e à previdência social são significativos, e por isso, o investimento em questões relacionadas a qualidade de vida não deve ser tratado como despesa.

Portanto, a importância do controle dos índices de absenteísmo está relacionada a estabilidade da empresa no mercado cada vez mais competitivo (JOURDAIN e CHÊNEVERT, 2015).

Em se tratando de controle dos índices de absenteísmo, uma vez que a identificação e a prevenção passam pela análise das informações geradas pelos empregados, é preciso registrá-las de forma abrangente como, por exemplo: registros do Código Internacional de Doenças (CID), exames médicos e fisioterapia (BUSTILLOS, VARGAS III e GOMERO-CUADRA, 2015).

Fazer previsões e identificar tendências tem sido desafiador para as empresas, por possuírem muitas informações armazenadas. Por essa razão, as técnicas de inteligência computacional surgem como alternativa de apoio ou substituição dos métodos tradicionais de análise para permitir a extração de informações valiosas em base de dados (LINOFF e BERRY,

2011).

Dessa forma, observa-se que o desenvolvimento e a aplicação de pesquisa fundamentada em técnicas de inteligência computacional na previsão do absentismo e identificação de tendências absenteístas, são relevantes na ampliação dos estudos sobre o assunto e, conseqüentemente para o tratamento desse comportamento oneroso para a empresa e prejudicial ao empregado (JOHNS, 2010).

A inteligência computacional também chamada de Computação Natural (CN) busca por meio de técnicas inspiradas na natureza, o desenvolvimento de sistemas inteligentes que imitem aspectos do comportamento humano, tais como: aprendizado, percepção, raciocínio, evolução e adaptação (ZURADA, MARKS e ROBINSON, 1995).

Técnicas de inteligência computacional podem ser aplicadas na área de pesquisa conhecida como Descoberta de Conhecimento em Bases de Dados, em inglês *Knowledge Discovery in Databases* (KDD). O KDD é um processo com fases que tem por objetivo extrair conhecimento de um conjunto de dados que possa ser aplicado na tomada de decisão (ZURADA e KARWOWSKI, 2011).

No processo do KDD, destaca-se a mineração de dados ou *Data Mining* (DM), conhecida como a etapa mais importante do KDD que contém o algoritmo minerador, responsável pela exploração e análise do conhecimento implícito contido em um banco de dados, com objetivo de descobrir padrões, regras e fornecer informações que permitam montar estratégias corporativas eficientes (LINOFF e BERRY, 2011; BRAHAMI, ATMANI e MATTA, 2013).

Diversas técnicas são utilizadas para a mineração de dados como as Redes Neurais Artificiais (RNAs) que são modelos constituídos por unidades de processamento simples, chamados de neurônios artificiais, que calculam funções matemáticas. Estes modelos são inspirados na estrutura do cérebro e têm como objetivo simular o comportamento humano, tais como: aprendizagem, associação, generalização e abstração quando submetidas a treinamento, são particularmente eficientes para o mapeamento entrada/saída de sistemas não lineares e para realizar processamento paralelo, além de simular sistemas complexos (HAYKIN, 2001).

Duas arquiteturas de RNAs dentre várias disponíveis podem ser aplicadas na solução de diversos problemas: a *MultiLayer Perceptron* (MLP) e a *Self-Organizing Map* ou rede SOM. Uma MLP consiste de um conjunto de neurônios, que constituem a camada de entrada, uma ou mais camadas ocultas e uma camada de saída, onde o sinal de entrada se propaga pela RNA camada por camada. Trata-se de uma RNA com aprendizagem supervisionada usada para

classificação (SILVA, SPATTI e FLAUZINO, 2010).

A *Self-Organizing Map* ou rede SOM é uma rede neural de aprendizagem competitiva organizada em duas camadas, a primeira camada representa o vetor dos dados de entrada, a segunda corresponde a uma grade de neurônios, geralmente bidimensional, totalmente conectada aos componentes do vetor de entrada. Trata-se de uma RNA com aprendizagem não supervisionada (HAYKIN, 2001). As principais aplicações da rede SOM estão no campo de agrupamento e visualização de dados. Estas características fazem com que a rede SOM seja muito utilizada na geração de agrupamentos (VESANTO e ALHONIEMI, 2000).

Prever e identificar tendências tem sido um desafio constante, a descoberta de conhecimento torna-se um dos recursos mais valiosos utilizados pelas empresas, ou seja, a previsão e identificação de tendências pode contribuir na tomada de decisão que são relevantes para reduzir perdas para a empresa e ao mesmo tempo à implantação de programas voltados a melhorar a qualidade de vida no trabalho do empregado.

1.1 Justificativa e motivação

O tema é relevante para as empresas porque incide diretamente em um problema organizacional silencioso e crônico, o absenteísmo reduz a produtividade na empresa de forma acentuada, reduzindo também a competitividade (LOHAUS e HABERMANN, 2019).

Considerável atenção tem sido direcionada ao fenômeno absenteísmo em virtude de seus impactos para o desempenho econômico e operacional das empresas, para os trabalhadores e para a sociedade, frequentes têm sido os estudos das causas e decorrências desses fenômenos (LOPES et al., 2017).

A taxa de absenteísmo é um importante indicador para mensurar as condições e qualidade de vida no trabalho, além de estar diretamente ligada ao lado financeiro das empresas (AGUIAR e OLIVEIRA, 2009; CUCCHIELLA, GASTALDI e RANIERI, 2014; LOPES et al., 2017).

Em se tratando de mudanças no comportamento, a sobrecarga física e a pressão no trabalho podem acarretar em sobrecarga psicológica aos empregados e contribuir para o desequilíbrio e ao estresse, que levam à deterioração da saúde mental manifestada principalmente pela depressão (FERREIRA e FERREIRA, 2015).

A depressão é considerada o mal do século XXI e também pode levar ao absenteísmo (HARGRAVE et al., 2008; BHUI et al., 2012). Segundo Collins (2017) a depressão aumentou

consideravelmente nas últimas décadas e com o uso de novas tecnologias incorporadas no dia a dia, também podem desencadear a depressão e conseqüentemente o absenteísmo.

Assim sendo, a previsão do absenteísmo e a identificação de tendências absenteístas são importantes e contribuem para reduzir perdas para a empresa e ao mesmo tempo melhorar a qualidade de vida do empregado. Contribui também para a tomada de decisão da empresa quanto à implantação de programas voltados a qualidade de vida no trabalho.

A importância do assunto levou autores a tratarem o absenteísmo com técnicas estatísticas como: Leão et al. (2015) em análise do perfil e os indicadores de absenteísmo-doença entre servidores municipais de Goiânia; Marques et al. (2015) em análise do absenteísmo da equipe de enfermagem de um hospital universitário; Johns (2015) em gerenciando o absenteísmo; Zhang, Sun e Woodcock (2017) em avaliação da perda de produtividade devido ao absenteísmo; Miranda, Araujo e Marcelino (2017) em análise da relação entre absenteísmo e desempenho acadêmico; Taylor (2018) em análise da resistência ao absenteísmo ao trabalho.

Lopes et al. (2017) realizaram estudos com modelo matemático *fuzzy*, porém relacionado ao fenômeno presenteísmo. Assim, técnicas de inteligência computacional ainda não foram aplicadas na previsão do absenteísmo e identificação de tendências absenteístas e, conjuntamente com a redução de perdas para a empresa e ao mesmo tempo melhorar a qualidade de vida do empregado, considera-se como lacuna de pesquisa e contribuição deste trabalho.

1.2 Problema de pesquisa

A velocidade com que a tecnologia evolui faz com que a cobrança para atingir os resultados seja cada vez maior, podendo refletir em estresse aos empregados ou provocar outros transtornos em seu estado de saúde, incluindo a depressão e causar absenteísmo.

Dessa forma, observa-se que o desenvolvimento e a aplicação de pesquisa fundamentada em técnicas de inteligência computacional na previsão do absenteísmo e identificação de tendências absenteístas, são relevantes na ampliação dos estudos sobre o assunto e, por conseguinte para o tratamento desse comportamento oneroso para a empresa e prejudicial ao empregado uma vez que a perda da qualidade de vida implica em problemas de saúde (JOHNS, 2008; JOHNS, 2010).

Verificou-se por meio de uma revisão bibliométrica que não existe aplicação de técnicas de inteligência computacional na previsão do absenteísmo e identificação de

tendências absenteístas. O trabalho mais próximo encontrado foi desenvolvido por Lopes et al. (2017) sobre o fenômeno presenteísmo usando um modelo matemático *fuzzy*.

Assim, o problema de pesquisa é explicitado a partir da seguinte indagação: Técnicas da inteligência computacional podem prever o absenteísmo e identificar tendências absenteístas?

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo geral deste trabalho foi aplicar técnicas da inteligência computacional na previsão do absenteísmo e na identificação de tendências absenteístas.

1.3.2 Objetivos específicos

Os objetivos específicos são:

- Geração da base de dados por meio da coleta, pré-processamento, enriquecimento e transformação dos atributos no período de dez anos.
- Seleção das técnicas de inteligência computacional para prever o absenteísmo e identificar tendências absenteístas.
- Previsão do absenteísmo com aplicação da MLP (Fase 1, Etapa 1) e dos *rough sets* com MLPs (Fase 1, Etapa 2).
- Comparação dos resultados computacionais da previsão do absenteísmo.
- Identificação de tendências absenteístas com a aplicação e avaliação da *Self-Organizing Map* (SOM) na geração de agrupamentos (Fase 2, Etapa 3).

1.4 Delimitação do tema de pesquisa

A empresa selecionada como objeto de estudo é uma empresa pública federal responsável pela execução do sistema de envio e entrega de correspondências no Brasil passou por diferentes fases desde a sua criação. Atualmente dispõe de inúmeros serviços de logística e produtos, está vinculada ao Ministério da Ciência, Tecnologia, Inovações e Comunicações (CORREIOS, 2019). Tem origem com a criação do Correio-mor das cartas do mar em 1º de fevereiro de 1663 (VALLE SALVINO, 2018).

A base de dados utilizada foi gerada ao longo de dez anos e gradativamente enriquecida com atributos relevantes para o estudo. As técnicas de inteligência computacional foram selecionadas devido à sua capacidade de apresentar tradicionalmente bons resultados na

classificação (RNA), redução de atributos (RS) e na formação de agrupamentos (SOM).

A previsão do absenteísmo e a identificação de tendências absenteístas foram escolhidas porque podem permitir que ações sejam tomadas antes da ocorrência do absenteísmo e por conseguinte, permitem o tratamento preventivo evitando sua ocorrência.

1.5 Organização do Trabalho

Esse trabalho foi estruturado em cinco capítulos. Além do capítulo de introdução (capítulo 1), o capítulo dois apresenta a fundamentação teórica, o capítulo três apresenta os materiais e métodos, o capítulo quatro apresenta os resultados dos experimentos computacionais e o capítulo cinco a conclusão.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo apresenta-se a fundamentação teórica sobre absenteísmo, descoberta de conhecimento em base de dados e as técnicas da inteligência computacional, além da revisão bibliométrica.

2.1 Absenteísmo

O absenteísmo é um fenômeno relacionado ao não comparecimento ao trabalho de forma frequente e habitual, o que geralmente indica que o trabalho é feito de forma menos eficiente ou que não é feito (JOHNS, 2010). Absenteísmo difere da falta ao trabalho em situação pontual isolada em função de sua frequência e habitualidade (JOHNS, 2008; PRATER e SMITH, 2011; LOHAUS e HABERMANN, 2019).

De acordo com Silva e Marziale (2012), o absenteísmo pode indicar insatisfação, desmotivação e sobrecarga da equipe de trabalho, o que pode revelar existência de problemas extremamente preocupantes, quando ocasionados por doença.

A conciliação da expectativa do empregado com as necessidades organizacionais, referenciada por Britt e Jex (2015), tornou-se um desafio diante das dificuldades apresentadas pelo ambiente de trabalho, como as situações de insalubridade, longas jornadas de trabalho, baixa remuneração, duplo emprego, tensão emocional e estresse, que podem ocasionar o absenteísmo.

Segundo Inoue et al. (2008) para que as empresas consigam identificar o real motivo das ausências dos seus empregados, faz-se necessário um estudo mais detalhado das causas e consequências do absenteísmo.

Hito (2007) enumerou as principais causas do absenteísmo: doença efetivamente comprovada e não comprovada, razões diversas de caráter familiar, atrasos involuntários ou por motivos de força maior, faltas voluntárias por motivos pessoais, dificuldades e problemas financeiros, problemas de transporte, desmotivação, supervisão precária da chefia, políticas inadequadas da empresa, jornadas prolongadas de trabalho, condições insalubres, tensão emocional e estresse.

Devido aos altos custos do absenteísmo, as empresas procuram administrá-lo de modo a reduzir ao máximo suas perdas. Esse tem sido um dos maiores desafios, por exemplo, para o setor de saúde ocupacional, dada a dimensão do fenômeno que engloba a detecção e o

monitoramento das causas de fundo psicológico, físicos e ambientais, considerando as circunstâncias que levam o empregado ao absenteísmo (ADDAE, JOHNS e BOIES, 2013; RUSSO et al., 2013).

É importante ressaltar que o nível de comportamento organizacional também pode desencadear doenças que levam ao absenteísmo, Bakker e Demerouti (2007) analisaram três níveis de comportamento organizacional:

- a) Nível individual - examina-se fundamentos da conduta pessoal: valores, atitudes, percepção, aprendizagem, o papel da personalidade e das emoções nesse comportamento, a motivação individual e o processo individual de tomada de decisão;
- b) Nível do grupo - pondera-se modelos de comportamento de grupo: temas relacionados à comunicação e à tomada de decisão dos grupos, liderança, confiança, poder, política, conflitos, negociação, dentre outros, e o
- c) Nível do sistema organizacional - discute-se como o comportamento pode ser afetado por diferentes dimensões das organizações. Além disso, aborda-se temas como as políticas e as práticas de recursos humanos e como elas influenciam os componentes das organizações e como a cultura institucional influencia o comportamento dos membros.

Estas influências dos níveis de comportamento organizacional derivam tanto de fontes diversas, quanto das influências dos gestores (KRISTENSEN et al., 2006), companheiros de trabalho (XIE e JOHNS, 2000; LIEKE et al., 2016), outros pares (BAMBERGER e BIRON, 2007) e a sociedade em geral (PATTON e JOHNS, 2007).

2.1.1 Estudos relacionados ao Absenteísmo

O interesse na causa do absenteísmo aumentou nas últimas três décadas devido ao alto custo para a empresa e a necessidade de melhoria da qualidade de vida no trabalho por parte dos empregados (MIRAGLIA e JOHNS, 2018).

O absenteísmo, visto como o fenômeno de não comparecer ao trabalho com frequência, tem sido amplamente pesquisado nas últimas décadas por causa de seu impacto na produtividade das empresas (LOHAUS e HABERMANN, 2019).

Avanços foram feitos nas pesquisas sobre o absenteísmo, em parte devido à grande diversidade de métodos de pesquisa que foram empregados para estudar o fenômeno. Os métodos melhoraram a compreensão sobre o absenteísmo no trabalho. Além disso, quase todos os indivíduos passaram pelo absenteísmo em algum momento, começando em uma idade jovem e continuando ao longo de suas vidas profissionais. A diversidade metodológica se expandiu

para uma ampla variedade de disciplinas acadêmicas e profissionais que se interessaram pelo absenteísmo (PATTON e JOHNS, 2012).

Oenning, Carvalho e Lima (2014) propuseram identificar fatores de risco ao absenteísmo com licença médica em trabalhadores de uma empresa de petróleo da Região Norte-Nordeste do Brasil entre 2007 e 2009, com todos os empregados da empresa. A variável resposta utilizada para representar o absenteísmo com licença médica foi a incidência média de faltas com licenças médicas no período e uma análise de regressão logística foi utilizada para investigar a associação entre incidência média de faltas > 5,0% no período. Os atributos utilizados pelos autores do trabalho são apresentados no Quadro 1.

Quadro1 – Atributos utilizados que identificam fatores de risco ao absenteísmo em trabalhadores de uma empresa de petróleo da Região Norte-Nordeste do Brasil entre 2007 e 2009

Variáveis/Atributos	Variáveis/Atributos
Sexo	Glicemia
Cargo	Diabetes não controlado
Idade	Doença do aparelho cardiovascular
Tempo de atuação	Digestivo
Regime de trabalho	Aparelho locomotor
Tabagismo	Neurológica
Hipertensão arterial	Neoplasia
Índice de massa corporal	Posturas forçadas no trabalho
Atividade física	Relacionamento com a chefia
Risco coronariano	Atenção concentrada no trabalho
Sono	

Fonte: Adaptado de Oenning, Carvalho e Lima (2014).

Os resultados mostraram que a incidência média de faltas com licenças médicas > 5,0% no período da corte foi 15,5%, 2,6 vezes mais chance de ser do sexo feminino, ser fumante ou ex-fumante, estar insatisfeito com o trabalho e relatar sono anormal foram preditores de absenteísmo ao trabalho por doença.

Laaksonen et al. (2010) apresentaram outros fatores associados ao absenteísmo. Tal pesquisa vê como propósito determinar quais arranjos de trabalho, condições físicas de trabalho e condições de trabalho psicossocial são importantes fatores de risco para a ausência de doença. Os dados da pesquisa foram vinculados aos registros de absenteísmo por doença do empregador para 3 anos subsequentes. As primeiras ocorrências de episódios de absenteísmo por doença a

curto prazo (1-3 dias), intermédios (4-14 dias) e a longo prazo (15 dias ou mais) foram examinadas pelo uso de modelos Bayesianos.

Os resultados mostraram que a carga de trabalho físico e as exposições perigosas foram consistentemente associadas a episódios de absenteísmo por doença maiores no curto, intermédios e longo prazo. Além disso, o baixo controle do trabalho nas mulheres e a insatisfação no trabalho nos homens foram consistentemente associados ao aumento do risco de absenteísmo por doença (LAAKSONEN et al., 2010).

Lund et al. (2006) trouxeram como fatores do absenteísmo o desconforto dos postos de trabalho, elevação ou transporte de cargas e o ato de empurrar ou puxar cargas, o objetivo do artigo foi examinar os efeitos do ambiente de trabalho físico na ausência de doença de longo prazo e investigar a interação entre fatores de riscos físicos e psicossociais entre 5.357 empregados na Dinamarca com o uso de análise de regressão.

Os resultados mostraram que o risco de início do absenteísmo por doença a longo prazo foi aumentado devido às posições de trabalho desconfortáveis, flexão ou torção extrema do pescoço ou das costas, principalmente quando o trabalho era realizado de pé ou agachando, levantando, transportando, empurrando e puxando cargas, aumentaram o risco de absenteísmo por doença de longo prazo (LUND et al., 2006).

O estudo de Lötters e Burdorf (2006) trouxeram como fator de absenteísmo a hipertensão arterial e a doença do sistema osteomuscular. O objetivo do estudo prospectivo de coorte com acompanhamento de 1 ano foi determinar fatores prognósticos para a duração do absenteísmo por doença devido a distúrbios musculoesqueléticos. Uma análise estatística foi feita com a regressão de risco proporcional de Cox com uma variável de interação com o tempo para cada fator de risco. Foram realizadas análises univariadas e multivariadas sobre distúrbios musculoesqueléticos e, separadamente, para dor lombar.

Os resultados apontaram que os principais fatores associados com o absenteísmo por doença mais longa foram: idade avançada, gênero, carga de trabalho percebida e saúde geral mais baixa para distúrbios do pescoço, ombros e membros superiores e incapacidade funcional, ciática, percepção do trabalhador sobre a capacidade de retorno ao trabalho, e queixas crônicas pela dor lombar (LÖTTERS e BURDORF, 2006).

Lahti et al. (2010) apontaram a atividade física como fator de redução de risco do absenteísmo por doença, quando os indivíduos eram vigorosamente ativos. A intenção dos autores foi examinar se o volume e a intensidade da atividade física estão associados a

subsequentes períodos de absenteísmo por doença de diferentes prazos, e quanto dessas associações podem ser explicadas pela posição socioeconômica, índice de massa corporal (IMC) e funcionamento da saúde física. Os dados foram coletados por questionários de 2000 a 2002 entre empregados de 40 a 60 anos da cidade de Helsinque na Finlândia com 6.465 empregados.

O volume de atividade física foi associado de forma fraca e um tanto incompatível com o absenteísmo por doença. No entanto, os empregados que estavam vigorosamente ativos sistematicamente reduziram o risco de absenteísmo por doença, enquanto o mesmo volume de atividade física moderadamente intensiva não reduziu o risco de absenteísmo por doença. (LAHTI et al., 2010).

Os resultados do artigo sugerem que a atividade física vigorosa está associada á redução do absenteísmo por doença e pode contribuir para uma melhorar a capacidade de trabalho (LAHTI et al., 2010).

Moreau et al. (2004) mostraram a obesidade associada a uma elevada incidência anual de absenteísmo por licença médica e a longos períodos de absenteísmo por doença, os autores enfatizaram que em muitos estudos, a obesidade foi associada a morbidade ou mortalidade, mas apenas algumas estudaram a relação entre obesidade e licença médica.

O estudo foi realizado em 25 empresas da Bélgica entre 1994 e 1998, com 20.463 empregados com idade entre 35 e 59 anos. O 3º percentil da distribuição dos dias anuais totais de doença foi utilizado como um corte para classificar os empregados com alta taxa de absenteísmo por doença. A relação entre absenteísmo por doença e obesidade e distribuição de gordura corporal avaliada pelo índice de massa corporal (IMC) e a circunferência da cintura, respectivamente, foi analisada por modelos de regressão logística multivariada.

Os resultados mostraram que a gordura abdominal central é um preditor independente de absenteísmo por doença em ambos os sexos com alta incidência de doença por longo prazo, mas não o IMC.

Addae, Johns e Boies (2013) enfatizaram que, apesar da grande quantidade de pesquisas sobre o absenteísmo, poucos estudos examinaram o fenômeno por uma perspectiva transcultural, ou seja, de uma perspectiva entre países e culturas. Com a evolução e o crescente interesse pela compreensão transcultural das atitudes, percepções e comportamentos dos empregados, os autores buscaram ampliar a pesquisa na área, investigando a percepção dos empregados quanto à legitimidade do absenteísmo em uma perspectiva transnacional. A

pesquisa sugere que ao fazê-lo, o estudo integrou conceitos previamente inexplorados no estudo do absenteísmo.

Miraglia e Johns (2018) elaboraram um modelo multinível da dinâmica social do fenômeno do absenteísmo que categoriza as várias forças do contexto social que moldam o comportamento dos empregados em sete níveis, indivíduo, grupo de trabalho, referência de vida, organização, ocupação, comunidade/localidade e nação/sociedade conforme mostra a Figura 1.

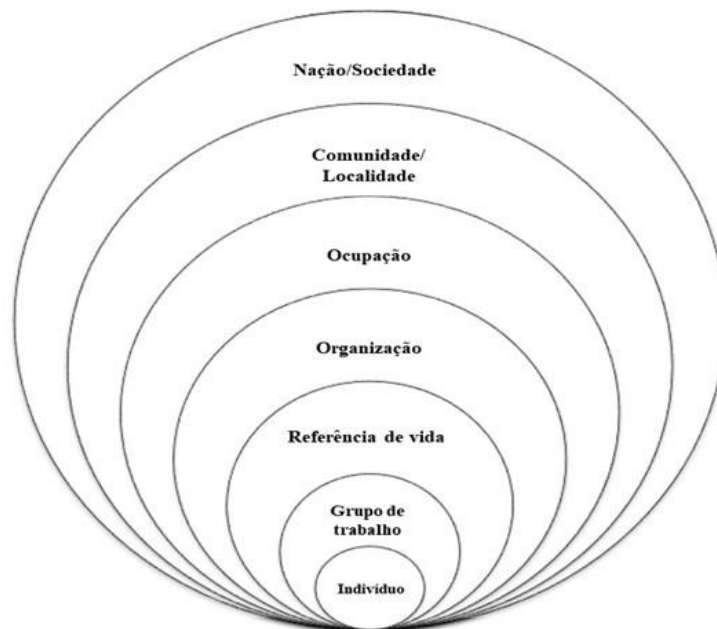


Figura 1 – Um modelo multinível da dinâmica social do fenômeno do absenteísmo. Adaptado de Miraglia e Johns (2018).

O artigo foi baseado em 291 trabalhos de grande relevância sobre o fenômeno do absenteísmo. Segundo os autores esse modelo multinível pode ser usado como um roteiro para entender quais são os atores sociais, dentro e fora da organização influenciam padrões e comportamentos de absenteísmo individuais e como essa influência social é exercida (MIRAGLIA E JOHNS, 2018).

Yu et al. (2019) realizaram um estudo recente sobre o estresse térmico que afeta a saúde dos trabalhadores e reduz a produtividade do trabalho causa absenteísmo, reduz a capacidade de trabalho e diminui a produtividade. O estresse térmico pode, portanto, ter efeitos econômicos, já que a saúde da força de trabalho é o insumo primário para a produção econômica. Além disso, o efeito do estresse térmico na produtividade do trabalho também é distribuído de forma desigual devido à diferente adaptabilidade do calor.

Lohaus e Habermann (2019) mostram que, até hoje, nenhuma definição uniforme nem

métodos consistentes de mensuração foram empregados na pesquisa sobre presenteísmo. Segundo os autores embora tenha havido várias tentativas de desenvolvimento da teoria, um modelo abrangente que pode explicar por que as pessoas escolhem presenteísmo ou absenteísmo ainda não foi estabelecido.

Existe outro fenômeno que merece ser considerado e que também afeta o trabalhador que se chama presenteísmo, definido como o fato dos empregados estarem no local de trabalho, mas, devido a problemas de ordem física ou psicológica, não cumprirem na totalidade as suas funções (HEMP, 2004; JOHNS, 2006; BÖCKERMAN e LAUKKANEN, 2009; BÖCKERMAN e LAUKKANEN, 2010).

O absenteísmo e o presenteísmo são vistos pelos pesquisadores como fenômenos relacionados (LOHAUS e HABERMANN, 2019). Segundo Fiorelli (2013), além das doenças relatadas em decorrência do estresse, pode-se observar outros fatores de diferentes origens que levam ao presenteísmo: frustração pelo fracasso na busca de um objetivo; conflito decorrente da competição entre motivações ou impulsos comportamentais incompatíveis; mudança resultante de uma alteração na forma de vida da pessoa; e pressão envolvendo expectativas ou exigências para que a pessoa se comporte de determinada forma. Esses fatores diferentes que levam ao presenteísmo podem desencadear doenças levando ao absenteísmo.

Macgregor, Cunningham e Caverley (2008) propuseram investigar a relação de eventos estressantes da vida e eventos relacionados à saúde com absenteísmo por doença e presenteísmo. Os achados indicaram que os eventos estressantes da vida foram significativamente associados com o absenteísmo e com o presenteísmo no mesmo grau, sugerindo que os empregados estão substituindo o presenteísmo pelo absenteísmo.

O estudo sugere que o absenteísmo e o presenteísmo são medidas importantes da saúde dos empregados e da produtividade organizacional.

2.2 Descoberta de Conhecimento em Bases de dados

Descoberta de Conhecimento em base de Dados ou *Knowledge Discovery in Databases* (KDD) é um processo de várias etapas não trivial, interativo e iterativo para identificação de padrões compreensíveis, válidos, e úteis a partir de grandes conjuntos de dados podendo ser aplicado em diversas áreas do conhecimento (BAGGA e SINGH, 2012).

O processo de KDD tem por objetivo descobrir novos padrões em bases de dados de maneira automática. Esse processo de descoberta é definido como a extração de conhecimento

de alto nível partindo de bases de dados reais e é composta de diversas fases (LAST, KLEIN e KANDEL, 2001).

O KDD utiliza diversas áreas para realizar os seus processos, tais como: métodos estatísticos, reconhecimento de padrões, visualização, banco de dados, aprendizado de máquina, inteligência artificial, *data warehouse*. É um processo composto por fases que devem ser desenvolvidas para atingir o objetivo final, que é extração de conhecimento (FAYYAD, PIATETSKY-SHAPIRO e SMITH, 1996).

2.2.1 Fases do KDD

As fases do KDD são:

- a) Seleção dos dados;
- b) Pré-processamento dos dados;
- c) Transformação dos dados;
- d) Mineração de dados e
- e) Interpretação/avaliação do conhecimento.

As três fases iniciais do KDD, a seleção, o pré-processamento e a transformação, são chamadas de preparação dos dados, são as fases que exigem mais tempo no processo devido a limpeza dos dados (RISTOSKI e PAULHEIM, 2016):

A Figura 2 apresenta as fases do processo de KDD.

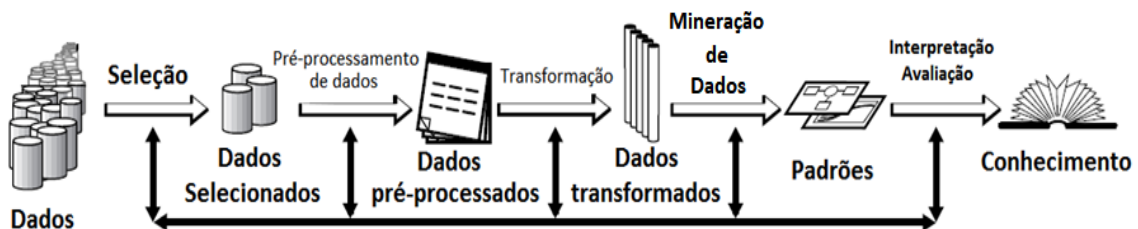


Figura 2 – Fases do processo de KDD. Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smith (1996).

O processo compreende todo o ciclo que o dado percorre até ser transformado em conhecimento. A iteração entre as fases pode ser observada pelas setas escuras.

As fases do KDD da Figura 2.1 são descritas a seguir:

- a) Seleção de Dados

A fase de seleção de dados é a primeira no processo de descobrimento de conhecimento. Nesta fase, busca-se determinar o subconjunto de atributos sobre o qual o KDD deverá atuar, facilitando o trabalho dos algoritmos responsáveis pela tarefa de organização e ordenação dos dados. Escolhem-se apenas atributos relevantes do conjunto de atributos da base de dados, que estejam de acordo com os objetivos da empresa. O subconjunto selecionado é então fornecido para o algoritmo de mineração dos dados até a transformação dos dados para torná-lo mais compreensíveis (RISTOSKI e PAULHEIM, 2016).

b) Pré-processamento e Limpeza dos Dados

Nesta etapa os dados redundantes e inconsistentes são identificados e podem ser excluídos, recuperam dados incompletos e avaliam possíveis dados discrepantes ao conjunto conhecido como *outliers* em busca da qualidade dos dados selecionados.

O auxílio do especialista do domínio é fundamental, pois é ele mesmo quem definirá se os padrões adquiridos são interessantes, se o conhecimento é válido, novo e útil, ou se será necessário retornar a alguma das etapas anteriores (ANUMALLA, 2007). O processo de limpeza dos dados também é chamado de eliminação de ruído ou redução de ruído e pode ser feito usando várias técnicas (PHRIDVIRAJ e GURURAO, 2014).

c) Transformação de Dados

A fase de transformação de dados ou codificação dos dados tem como objetivo principal converter o conjunto bruto de dados em uma forma padrão de uso, ou seja, os dados do conjunto bruto devem ser codificados para que possam ser usados como entrada dos algoritmos de mineração de dados (LINOFF e BERRY, 2011).

Alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. É necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos. Não há um critério específico para a transformação dos dados e diversas técnicas podem utilizadas conforme os objetivos, os quais seguem: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores mais genéricos), normalização (coloca as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes) (HAN, KAMBER e PEI, 2011).

d) Mineração de Dados

Todas as etapas do processo de KDD tem sua importância para o sucesso do mesmo. Entretanto, é a etapa da mineração de dados que recebe o maior destaque na literatura. De acordo com Jothi, Rashid e Husain (2015), a mineração de dados, é considerada a etapa mais importante do processo pela existência do algoritmo minerador, que diante da tarefa especificada será capaz de extrair de modo eficaz o conhecimento implícito e útil de um banco de dados.

A expressão mineração de dados ou em inglês *Data Mining*, surgiu devido a algo em comum entre a procura de informação que seja importante numa base de dados e o ato de minerar a montanha em busca de um veio de ouro (TURBAN, et al., 2010). A mineração de dados é o elemento responsável pela extração eficiente do conhecimento implícito e útil contido em um banco de dados (BIGUS, 1996).

Durante a etapa de mineração de dados é realizada a busca concreta por conhecimento útil no contexto da aplicação do KDD. Na qual são definidos as técnicas e os algoritmos a serem utilizados no problema em questão. A escolha da técnica depende, muitas vezes, do tipo de tarefa do KDD a ser realizada (PIATETSKY-SHAPIRO, MATHEUS e CHAN, 1993; LINOFF e BERRY, 2011; SILVA, PERES e BOSCARIOLI, 2016).

Camilo e Silva (2009) relacionaram algumas técnicas utilizadas na mineração de dados: Árvores de Decisão, Classificação Bayesiana, Redes Neurais Artificiais (*Artificial Neural Networks*), *Support Vector Machines* (SVM), Algoritmo Genético (AG), Teoria dos Conjuntos Aproximados (*Rough Sets Theory*), Teoria dos Conjuntos *Fuzzy*, *k-Means* e Métodos Baseados na Densidade (*Density-Based Methods*).

A essência do conhecimento extraído deve ser considerada um quesito imprescindível da mineração de dados e deve-se levar em conta a precisão, a compreensibilidade, a surpresa e/ou o eventual interesse pelo conhecimento adquirido (POMBO, ARAÚJO e VIANA 2014).

A Figura 3 mostra um esquema simplificado da mineração de dados.



Figura 3 – Esquema simplificado da mineração de dados. Fonte: Adaptado de Sassi, Silva e Hernandez (2008).

Esse processo inicia-se com os dados, os quais em seguida são transformados em informações válidas, isto é, o conjunto de dados é codificado. Após essas etapas, os dados são interpretados, coletados e transformados em conhecimento através da mineração de dados para auxiliar nas decisões e no crescimento organizacional (COSTA e ANDRADE NETTO, 2007).

e) Interpretação/Avaliação do conhecimento

A Interpretação/Avaliação do conhecimento também é conhecida como pós-processamento, com o intuito de garantir um bom grau de compreensão do conhecimento descoberto pelo algoritmo minerador, validando-o através de medidas da qualidade da solução e da percepção do analista de dados em medir o quão novo é o conhecimento descoberto, esta etapa também pode envolver a visualização dos padrões e os modelos extraídos, ou a visualização dos dados usando os modelos extraídos (RISTOSKI e PAULHEIM, 2016).

2.2.2 Tarefas do KDD

Existem várias formas de interpretação dos dados pelo KDD denominadas tarefas. Diversas tarefas podem ser realizadas, as mais comuns são associação, classificação, agrupamento e visualização descritos a seguir (FAYYAD, PIATETSKY-SHAPIRO e SMITH, 1996):

a) Regras de Associação

De acordo com Goldschmidt, Passos e Bezerra (2015), as regras de associação fundamentam-se na ação de encontrar um grupo de itens semelhantes, que ocorrem simultaneamente e de forma repetida em uma base de dados.

Estas regras descobertas devem considerar importantes não apenas associações triviais, mas também aquelas não óbvias onde jamais se imaginaria haver uma relação, das quais se transformam em importante fonte de informação na tomada de decisão (WITTEN et al., 2016).

Uma regra de associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (SADATH, 2013).

O objetivo refere-se à possibilidade de encontrar tendências a partir de regras de associação extraídas de bases de dados, podendo ser incrementada estrategicamente nas vendas de um determinado segmento comercial, como promoções entre produtos, rearranjo da disposição dos produtos em prateleiras e gôndolas, entre outras (THANUJA, VENKATESWARLU e ANJANEYULU, 2011).

b) Classificação

A tarefa de classificação corresponde em encontrar uma função que associe um caso a uma classe dentro de diversas classes discretas. Esta tarefa pode ser definida como uma função de aprendizado, que mapeia itens de dados em uma das diversas classes previamente definidas (KOTSIANTIS, ZAHARAKIS e PINTELAS, 2007).

Dessa forma a classificação consiste em obter um modelo baseado em um conjunto de exemplos que descrevem uma função não-conhecida. Neste tipo de tarefa, o objetivo é descobrir um relacionamento entre um atributo meta e um conjunto de atributos previsores, de modo que o processo de classificação possa usar esse relacionamento para prever a classe de um exemplo novo e desconhecido (FAYYAD, PIATETSKY-SHAPIRO e SMITH, 1996).

Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, contendo a indicação à qual classe pertence, a fim de aprender como classificar um novo registro, ou seja, o aprendizado é supervisionado (NASCIMENTO JUNIOR e YONEYAMA, 2004).

A classificação pode ser utilizada com êxito, por exemplo, em um conjunto de dados de um banco, em que se deseja descobrir qual o perfil dos clientes que aplicam seu dinheiro em fundos de investimento de longo prazo. Com classificadores pode-se prever que clientes do sexo masculino, com renda superior a R\$ 5000,00 e com idade acima de 30 anos apliquem em fundos de investimento de longo prazo. Neste caso, o atributo aplica em fundos de investimento de longo prazo é denominado classe, pois é o atributo alvo da classificação cujos possíveis valores, neste caso, são sim ou não (SADATH, 2013).

c) Análise de Agrupamento

O agrupamento difere da classificação, pois a primeira, visa criar os grupos através da organização dos elementos, enquanto a segunda, procura alocar elementos em classes já pré-definidas. O agrupamento transforma registros com grande número de atributos em conjuntos relativamente menores (BRAGA, CARVALHO e LUDERMIR, 2011).

Segundo Berkhin (2006), o agrupamento é uma tarefa onde se procura determinar um conjunto finito de conjuntos ou agrupamentos para descrever os dados, consiste na identificação de grupos semelhantes de objetos onde cada grupo é uma classe.

A tarefa de análise de agrupamento difere da classificação, pois não necessita que os registros sejam previamente categorizados, ou seja, o aprendizado é não-supervisionado. No

aprendizado não supervisionado, não há um professor para supervisionar o processo de aprendizagem (HAYKIN, 2001).

Dentro da mesma classe os objetos são semelhantes e se comparado com as outras classes são divergentes. Essa tarefa é usada para particionar os registros de uma base de dados, de forma que elementos em um agrupamento compartilhem um conjunto de propriedades comuns que os distinguem dos elementos de outros agrupamentos com o objetivo de maximizar similaridade intra-*cluster* e minimizar similaridade inter-*cluster* (KOHONEN, 2001).

d) Visualização

A tarefa de visualização é um processo para transformar informação em uma forma visual, permitindo aos usuários observar a informação, que pode ser usada para dar suporte ao processo de decisão quando as técnicas de mineração de dados requerem grande interação com o usuário de forma complexa (VESANTO e ALHONIEMI, 2000).

Esta tarefa tem sido usada para analisar e mostrar grandes volumes de dados multidimensionais. A visualização permite representar, diferentemente dos métodos estatísticos, os resultados sem necessariamente saber que tipo de fenômeno deve ser analisado (SHIRAVI, SHIRAVI e GHORBANI, 2012).

No caso de técnicas de agrupamento, podem ser usadas ferramentas de visualização para determinar qual, ou quais agrupamentos criados são úteis ou interessantes para os métodos de mineração de dados.

As ferramentas de visualização podem ser usadas para auxiliar, ou serem auxiliadas, pelas técnicas de mineração de dados. Por meio dessas ferramentas podem ser encontradas características, ou fenômenos, pouco comuns ou interessantes sem que se esteja diretamente procurando por eles (SILVA NETO et al., 2010).

A visualização de dados quando resulta da mineração de dados deve ser facilmente compreendida e diretamente utilizável, isso exige que as representações do conhecimento sejam expressivas, com interfaces de fácil utilização e com representações gráficas de visualização variadas (WU et al., 2010).

As representações gráficas ajudam na visualização dos dados, contribuindo com o estudo primário das características dos dados, dessa forma auxiliando na tarefa de mineração de dados. Evidenciam intervalos de maior e menor representatividade, média, mediana, limite inferior, limite superior, faixa interquartil, assimetria, simetria e *outlier*, que podem ajudar na interpretação inicial dos dados (LI et al., 2016).

A visualização de dados facilita a compreensão, e por esse motivo é bastante utilizada na mineração de dados. Existem diversos tipos de representações gráficas para visualização de dados que podem ser encontrados no trabalho de (WITTEN et al., 2016).

2.3 Técnicas da Inteligência Computacional

A inteligência computacional busca por meio de técnicas inspiradas na natureza, o desenvolvimento de sistemas inteligentes que imitem aspectos do comportamento humano, tais como: aprendizado, percepção, raciocínio, evolução e adaptação (ZURADA, MARKS e ROBINSON, 1995). Envolve ideias, pontos de vista, conceitos e técnicas de diversas áreas, dentre as quais podemos citar: Filosofia, Matemática, Economia, Neurociência, Psicologia, Linguística, Ciência da Computação (GOLDSCHMIDT, 2010).

A inteligência computacional tem aplicações em diferentes áreas do conhecimento como, por exemplo: detecção antecipada de eventos de inundação (FOTOVATIKHAH et al., 2018), processamento de imagens por sensoriamento remoto (ZHONG et al., 2018), previsão de carga de energia (FALLAH et al., 2018), modelagem da produção de hidrogênio (ARDABILI et al., 2018), *big data* (VIDAL-GARCÍA, VIDAL e BARROS, 2019), aprendizagem de máquina (MIIKKULAINEN et al., 2019), otimização (DHIMAN e KAUR, 2019).

2.3.1 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são modelos constituídos por unidades de processamento simples, chamados de neurônios artificiais, que calculam funções matemáticas. Esses modelos são inspirados na estrutura do cérebro e têm como objetivo simular o comportamento humano, tais como: aprendizagem, associação, generalização e abstração quando submetidas a treinamento. Uma importante característica das RNAs é a sua capacidade de aprender a partir de dados incompletos e sujeitos a ruídos (HAYKIN, 2001).

Os neurônios artificiais simulam as células neurais biológicas de uma maneira geral. As sinapses são representadas por um peso que é combinado linearmente com a entrada do neurônio e depois enviado ao próximo neurônio para fornecer um padrão de ativação. Semelhante aos neurônios biológicos, é necessário atingir um certo nível de ativação,

excedendo um certo limite (limiar) para que um neurônio dispare (KAPANOVA, DIMOV e SELLIER, 2016).

Para compreender a lógica de funcionamento das redes neurais artificiais, alguns conceitos básicos referentes ao funcionamento do cérebro humano e seus componentes, os neurônios, são de fundamental importância. A Figura 4 apresenta de forma simplificada um neurônio biológico e um neurônio artificial.

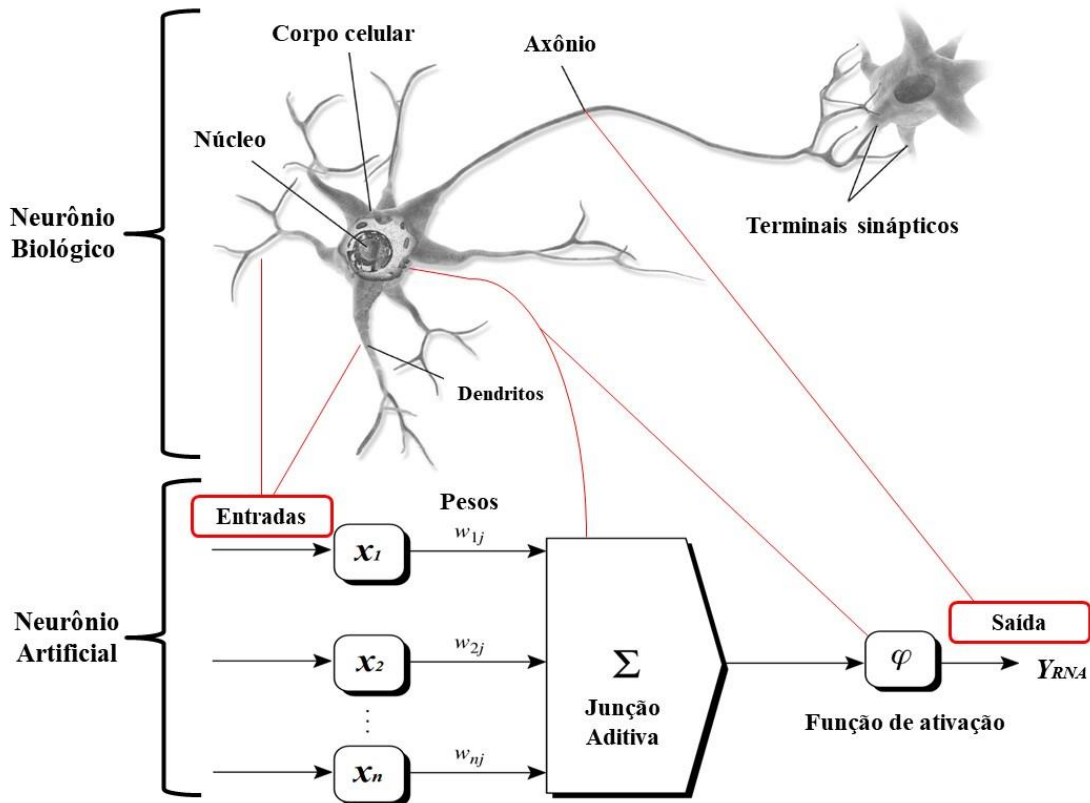


Figura 4 – Representação simplificada de um neurônio biológico e um neurônio artificial. Fonte: Adaptado de Haykin (2009); Kapanova, Dimov e Sellier (2016); Heidari e Shamsi (2018).

Os neurônios biológicos são divididos de maneira simplificada: o corpo celular também chamado de soma, responsável por coletar e combinar informações vindas de outros neurônios, os dendritos recebem as informações, ou impulsos nervosos, oriundas de outros neurônios e as conduzem até o corpo celular e o axônio constituído de uma fibra tubular que pode alcançar até alguns metros e é responsável por transmitir os estímulos para outras células. Cada um com funções específicas, porém complementares (BRAGA, CARVALHO e LUDERMIR, 2011).

A informação é processada no corpo celular e novos impulsos são gerados. Esses impulsos são transmitidos a outros neurônios, passando por meio do axônio até os dendritos

dos neurônios seguintes. O ponto de ligação entre terminação axônica de um neurônio e o dendrito de outro neurônio é chamado de terminais sinápticos. É pelos terminais sinápticos que os neurônios se unem funcionalmente formando as redes neurais biológicas. As sinapses controlam a transmissão de impulsos, ou seja, controlam o fluxo da informação entre os neurônios da rede neural biológica (BRAGA, CARVALHO e LUDERMIR, 2011).

Nas redes neurais artificiais, a aprendizagem se dá por meio de um conjunto de unidades simples de processamento chamados de neurônios artificiais ou nós divididos da seguinte maneira: os dados ou vetores de dados de entrada da rede (x_1, x_2, \dots, x_n) , os neurônios da camada de entrada da rede $(w_{1j}, w_{2j}, \dots, w_{nj})$ com seus respectivos pesos são transmitidos para a junção aditiva ou soma representada pela letra sigma, em seguida observa-se a função de ativação do neurônio artificial e finalmente a camada de saída (Y_{RNA}) , formada por um neurônio.

Uma importante característica das RNAs é a sua capacidade de aprender a partir de dados incompletos e sujeitos a ruídos. Em um sistema computacional convencional, se uma parte falha, em geral, o sistema como um todo se deteriora, enquanto que, em uma RNA, a tolerância às falhas faz parte da arquitetura, devido à sua natureza distribuída de processamento. Se um neurônio falha, sua saída errônea é sobrescrita pelas saídas corretas de seus elementos vizinhos. Assim, a princípio, uma RNA exhibe uma degradação suave do desempenho em vez de apresentar uma falha catastrófica (HAYKIN, 2009).

A ativação do neurônio artificial é realizada através da função de ativação, a qual desempenha tarefa semelhante à sinapse no neurônio biológico, transmitindo ou bloqueando os impulsos nervosos. De uma forma geral, o aprendizado das RNAs se dá através dos ajustes dos pesos sinápticos. O valor do peso sináptico $W^{(t+1)}$ no instante $t+1$, será determinado em função do seu valor na iteração anterior w^t , conforme a Equação (1):

$$w_i^{t+1} = w_i^t + \Delta w_i^t \quad (1)$$

A atualização dos pesos depende do algoritmo, mas geralmente baseia-se na minimização do erro \mathcal{E}_i , entre os valores previstos pela rede e as saídas y_i desejadas, conforme Equação (2):

$$\mathcal{E}_i = \sum w_i x_i - y_i \quad (2)$$

Desta forma, o aprendizado em uma RNA é definido como o ajuste iterativo dos pesos sinápticos, de forma a minimizar os erros, uma definição geral do que vem a ser aprendizado em uma RNA pode ser expressa da seguinte maneira: aprendizado é o processo

pelo qual os parâmetros de uma RNA são ajustados através de uma forma continuada de estímulo pelo ambiente no qual a rede está operando, sendo o tipo específico de aprendizagem realizada definido pela maneira particular como ocorrem os ajustes realizados nos parâmetros (KAPANOVA, DIMOV e SELLIER, 2016).

Diversos métodos para aprendizado foram desenvolvidos, podendo ser agrupados em dois paradigmas principais: aprendizado supervisionado e aprendizado não-supervisionado. No aprendizado supervisionado, existe um conhecimento prévio sobre os valores das entradas x_i e respectivas saídas y_i . A este conjunto de pares ordenados (x_i, y_i) , que se conhece a priori, dá-se o nome de base de dados de aprendizado. O algoritmo mais difundido é o de retropropagação do erro (*error back-propagation*) utilizado pela RNA do tipo *Multilayer Perceptron* (MLP).

O algoritmo de treinamento *error back-propagation* funciona da seguinte maneira: apresenta-se um padrão à camada de entrada da rede, este padrão é processado, camada por camada, até que a saída forneça a resposta processada, f_{MLP} , calculada como mostrado na Equação (3). Onde v_l e w_{lj} são pesos sinápticos; b_{l0} e b_0 são os *biases* e φ a função de ativação.

$$f_{MLP}(x) = \varphi \left(\sum_1^{Non} v_l \cdot \varphi \left(\sum w_{lj} x_l + b_{l0} \right) + b_0 \right) \quad (3)$$

Na aprendizagem não supervisionada não se conhece os valores das saídas desejadas y_i , então a aprendizagem ocorre através da identificação de padrões nas entradas. O aprendizado de uma RNA, na maioria dos casos, acontece com um subconjunto de exemplos (vetores de dados) que definem o chamado conjunto de treinamento e o teste da RNA é realizado com outro subconjunto de exemplos (vetores de dados) que definem o chamado conjunto de teste (HAYKIN, 2001).

A arquitetura de uma RNA depende do tipo de problema no qual a rede deve ser utilizada, pois ela é definida, entre outros fatores, pelo número de camadas, número de nós em cada camada, pelo tipo de conexão entre os nós (*feedforward* ou *feedback*) e por sua topologia. As RNAs podem ser treinadas utilizando-se valores iniciais aleatórios para as conexões de pesos. Os parâmetros de aprendizado são inicializados e os padrões de treinamento dos vetores de dados são apresentados para a RNA. Ao longo do progresso do treinamento são ajustadas as conexões de pesos e é possível monitorar o desempenho da RNA a cada época (HAYKIN, 2001).

As RNAs podem ser usadas quando há pouco conhecimento das relações entre atributos e classes, são adequadas para entradas e saídas de valor contínuo, ao contrário da maioria dos algoritmos, são bem-sucedidas em uma grande variedade de problemas do mundo real, incluindo o reconhecimento de caracteres manuscritos, patologias e medicina (HAYKIN, 2001).

2.3.1.1 *Multilayer Perceptron*

Uma *Multilayer Perceptron* (MLP) consiste de um conjunto de unidades (nós ou neurônios), que constituem a camada de entrada, uma ou mais camadas ocultas e uma camada de saída, onde o sinal de entrada se propaga pela RNA camada por camada. Na Figura 5 é apresentada a estrutura básica de uma MLP.

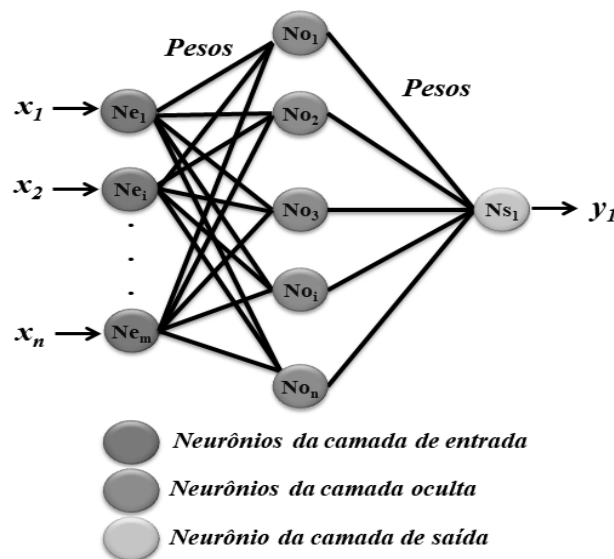


Figura 5 – Estrutura básica da MLP. Fonte: Adaptado de Sassi (2006).

Na Figura 5 pode-se observar os dados (vetores de dados) de entrada da rede (x_1, \dots, x_n), os neurônios da camada de entrada da rede (Ne_1, \dots, Ne_m) com seus respectivos pesos, os neurônios que formam a camada intermediária da rede (No_1, \dots, No_n) e a camada de saída (Ns_1), formada por um neurônio.

A MLP possui a capacidade de aprender por exemplos e fazer interpolações e extrapolações do que aprenderam. Um conjunto de procedimentos bem definidos, para adaptar os pesos de uma MLP para que ela possa aprender uma determinada função, é chamado algoritmo de treinamento ou de aprendizado (BIGUS, 1996).

O aprendizado de uma MLP usa um conjunto de dados correspondentes a uma amostra de sinais para entrada e saída do sistema. Para esse treinamento, a rede utiliza algoritmos de aprendizado.

Inicialmente a rede permanece inerte e o algoritmo de aprendizado modifica individualmente os pesos das interconexões de tal forma que o comportamento da rede reflita a ação desejada. Em outras palavras, a rede pode alterar sua estrutura interna de maneira incremental até que se alcance o desempenho esperado de estimação dos dados (SIMÕES e SHAW, 2007).

Como resultado do treinamento, a MLP produzirá valores de saída similares ao conjunto de dados para valores que sejam iguais às amostras de treinamento. Para valores intermediários, a rede produzirá uma interpolação. Ou seja, a MLP pode aprender através de exemplos (SIMÕES e SHAW, 2007).

O algoritmo de retroprogramação do erro (*error back-propagation*), utilizado na MLP, consiste, basicamente, em determinar as variações nos pesos sinápticos da MLP, tendo como objetivo minimizar o erro obtido na saída através do aprendizado do vetor de treinamento (entrada-saída) (RUMELHART et al., 1986).

Para isso, o algoritmo baseia-se no método do gradiente descendente, o qual, dada uma medida do erro, procura modificar o conjunto de pesos w_{ij} da rede, reduzindo o erro na direção mais íngreme da superfície definida no espaço w (HAYKIN, 2001).

Em resumo, o algoritmo do gradiente descendente estabelece mudanças nos pesos w_{ij} por uma quantidade Δw_{ij} proporcional ao gradiente do erro.

O algoritmo de retroprogramação do erro (*error back-propagation*) funciona da seguinte forma: apresenta-se um padrão à camada de entrada da rede, esse padrão é processado camada por camada até que a camada de saída forneça a resposta processada, f_{MLP} , como mostra a Equação (4):

$$f_{MLP}(x) = \varphi \left(\sum_1^{Non} v_l \cdot \varphi \left(\sum w_{ly} x_l + b_{l0} \right) + b_0 \right) \quad (4)$$

Na qual v_l e w_{lj} são pesos sinápticos; b_{l0} e b_0 são os biases; e φ a função de ativação, comumente especificada como sendo a função sigmoide, conforme a Figura 6.

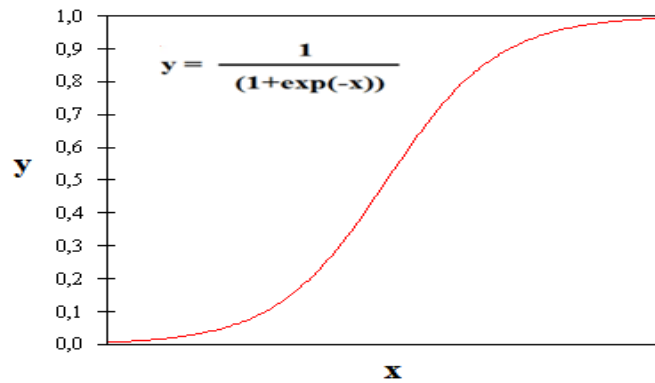


Figura 6 – Função sigmoide. Fonte: Adaptado de Sassi, Silva e Hernandez (2008).

O princípio deste algoritmo, para cálculo dos erros nas camadas intermediárias, é o seu cálculo por retroalimentação, possibilitando, desta forma, o ajuste dos pesos proporcionalmente aos valores das conexões entre camadas. De acordo com Haykin (2001), a MLP possui as seguintes características: função de ativação não-linear (sigmoide), uma ou mais camadas de neurônios ocultos e um alto grau de conectividade. Uma MLP treinada com o algoritmo de retropropagação realiza um mapeamento não-linear de entrada-saída. Ainda de acordo com Haykin (2001), matematicamente, uma MLP é um método para determinar uma função $f: A \rightarrow B$ onde A é um conjunto de dados, sendo que alguns elementos de B são conhecidos a priori.

De acordo com Simões e Shaw (2007), em um processo iterativo como o algoritmo *error back-propagation*, uma questão importante é sobre o critério de parada do treinamento. Considera-se em termos absolutos, o índice de desempenho (erro global) igual a zero (0) ou um (1) como valores considerados pequenos. Contudo, se durante a fase de treinamento o erro não for pequeno e não tiver tendência a diminuir, dois enfoques podem ser adotados:

- a) limitar o número de iterações, que consiste em encerrar o treinamento após o erro quadrático médio ficar abaixo de um valor a pré-definido. Este valor depende muito do problema. Haykin (2001) e Zurada (1992) sugerem estabelecer um valor elevado a 10^{-2} .

- b) amostrar e tirar a média de certo número de épocas, por exemplo, 100 épocas. Se o erro médio do último conjunto de 100 épocas não for melhor que o das 100 anteriores o treinamento deve ser cessado, indicando que um mínimo local foi alcançado. Depois disso, a rede deve estar ainda pronta para ser reutilizada para testes (HAYKIN, 2001).

Se o desempenho não for satisfatório, aos pesos devem ser adicionados um pouco de ruído branco para auxiliar a rede a sair do mínimo local. Ou então, a rede deve ser

completamente reinicializada e o processo de treinamento deve ser retomado (SIMÕES e SHAW, 2007).

2.3.2 *Self-Organizing Map*

A rede SOM (*Self-Organizing Map*) ou Mapa Auto Organizável é uma rede neural artificial, baseada em aprendizado competitivo e não supervisionado, que define um mapeamento de um espaço de entrada contínuo para um conjunto discreto de vetores de referência, os quais são organizados em uma camada uni ou bidimensional (KOHONEN, 2001).

Cada neurônio tem a mesma dimensão do espaço de entrada, e para cada padrão de entrada, um neurônio é definido como vencedor, para tanto, utiliza-se o critério de maior similaridade. O mapa é organizado de maneira que os neurônios próximos no mapa representem o mesmo agrupamento, uma vez que o ajuste dos pesos do neurônio vencedor e seus vizinhos é tal que faz com que o vetor de pesos se aproxime do vetor apresentado na entrada (LAINE, 2002).

Os neurônios da rede SOM estão localmente interconectados por uma relação de vizinhança, determinando a topologia do mapa. Num mapa bidimensional, a vizinhança pode ser triangular, retangular ou hexagonal (KOHONEN, 2001). Os neurônios se tornam seletivamente sintonizados a vários padrões de entrada ou classes de padrões de entrada no decorrer de um processo de aprendizado. O formato do arranjo ou topologia influencia diretamente na adaptação da rede SOM (HAYKIN, 2001; KOHONEN, 2001).

A Figura 7 ilustra uma grade (a) unidimensional retangular, (b) unidimensional hexagonal, (c) bidimensional retangular e (d) bidimensional hexagonal.

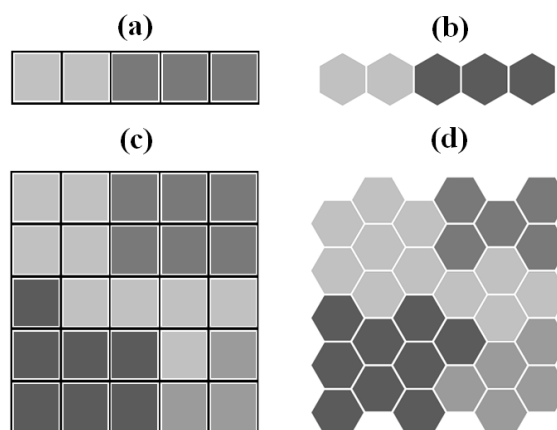


Figura 7 – Mapa unidimensional e bidimensional retangular e hexagonal. Fonte: Adaptado de Breard (2017).

Mapas de dimensionalidade mais alta são também possíveis, mas não são tão comuns. Os neurônios se tornam seletivamente sintonizados a vários padrões de entrada ou classes de padrões de entrada no decorrer de um processo de aprendizagem (HAYKIN, 2001).

O algoritmo responsável pela formação do Mapa Auto Organizável começa inicializando os pesos sinápticos da grade. Isto pode ser feito atribuindo-lhes valores pequenos tomados de um gerador de números aleatórios, fazendo dessa forma, nenhuma organização prévia imposta ao mapa de características (HAYKIN, 2001).

2.3.3 O Algoritmo de Aprendizado da Rede SOM

O algoritmo de aprendizado da rede SOM é composto por três fases: competição, cooperação e adaptação:

a) Competição. Para cada padrão de entrada, os neurônios da grade calculam seus respectivos valores de uma função discriminante. Esta função discriminante fornece a base para a competição entre os neurônios. O neurônio com o maior valor da função discriminante é declarado vencedor da competição (HAYKIN, 2001).

O princípio para o processo de aprendizado competitivo é a concorrência entre os neurônios, com o objetivo de determinar o vencedor da competição, pois o processo é não supervisionado, ou seja, não tem a saída desejada (KOHONEN, 2001).

O prêmio para o neurônio vencedor da competição é o ajuste dos seus pesos, proporcionalmente aos valores do padrão de entrada apresentado, visando melhorar o seu estado para a próxima competição, se todas as conexões laterais deste neurônio vencedor forem nulas, implica que somente os seus pesos serão ajustados. Este tipo de inibição mútua também é conhecido como *winner takes all* ou o vencedor leva tudo (KOHONEN, 2001; HAYKIN, 2001; BRAGA, CARVALHO e LUDERMIR, 2011; KOHONEN, 2013).

b) Cooperação. O neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados, fornecendo assim a base para a cooperação entre os neurônios vizinhos.

c) Adaptação Sináptica. Este último mecanismo permite que os neurônios excitados aumentem seus valores individuais da função discriminante em relação ao padrão de entrada através de ajustes adequados aplicados a seus pesos sinápticos. Os ajustes feitos são tais

que a resposta do neurônio vencedor à aplicação subsequente de um padrão de entrada similar é melhorada (HAYKIN, 2001).

Para cada dado apresentado à rede haverá uma competição entre todos os neurônios pelo direito de representá-lo. Vence a competição o neurônio que tiver o vetor de pesos mais próximo do vetor de dados (HAYKIN, 2001). A Figura 8 mostra uma grade bidimensional de neurônios da rede SOM.

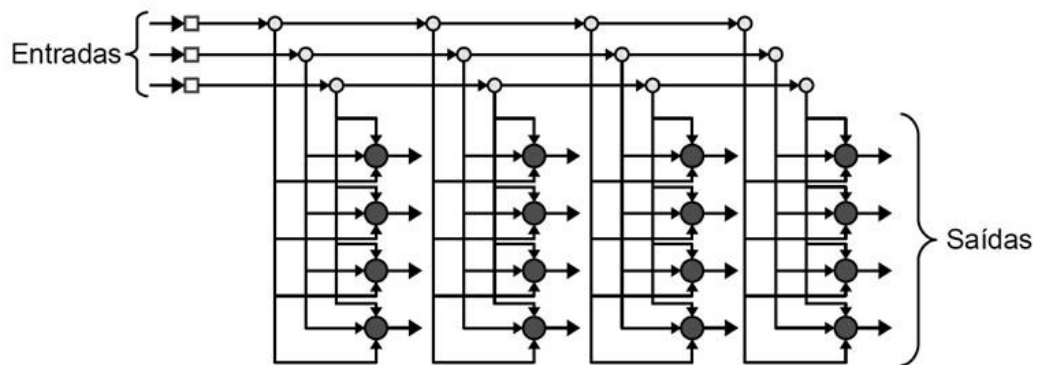


Figura 8 – Grade bidimensional de neurônios da rede SOM. Fonte: Adaptado de Jarske, Seabra e Silva (2018).

O algoritmo da rede SOM utiliza o chamado aprendizado competitivo. Uma rede neural com aprendizado competitivo é uma rede de uma única ou duas camadas (uni ou bidimensional) em que todos os neurônios recebem a mesma entrada. Cada neurônio computa o seu nível de ativação multiplicando o seu vetor de pesos pelo vetor de dados da entrada. O neurônio que tiver o maior nível de ativação é chamado de neurônio vencedor ou *Best Match Unit* (BMU), ou seja, o padrão de entrada que estiver sendo apresentado à rede provocará a ativação de apenas um neurônio de saída ou um neurônio por grupo (KOHONEN, 2001; DEMIR e CERGIBOZAN, 2018).

2.3.4 O Processo Competitivo da rede SOM

Para descrever os passos envolvidos com o processo de aprendizado competitivo utilizado na rede SOM, considera-se uma estrutura neural constituída de apenas uma camada neural (SILVA, SPATTI e FLAUZINO, 2010). A Figura 9 mostra um modelo da rede SOM.

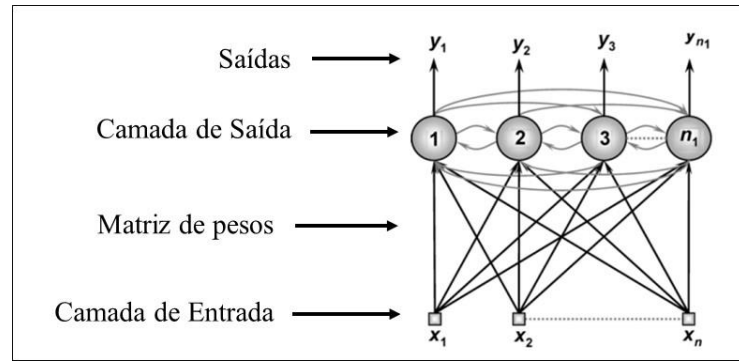


Figura 9 – Modelo da rede SOM. Fonte: Adaptado de Silva, Spatti e Flauzino (2010).

As conexões laterais entre os neurônios da Figura 9 assumem o papel de que um neurônio pode influenciar na resposta de saída produzida por outro neurônio.

Conforme Haykin (2001), se m representa a dimensão do espaço de entrada de dados, considere que um padrão vetor de entrada selecionado aleatoriamente do espaço de entrada seja representado pela Equação (5).

$$x = [x_1, x_2, \dots, x_m]^T \quad (5)$$

O vetor de peso sináptico de cada neurônio da grade tem a mesma dimensão que o espaço de entrada. Suponha que o vetor de peso sináptico do neurônio j seja representado pela Equação (6),

$$W_j = [w_{j1}, w_{j2}, \dots, w_{jm}]^T, j = 1, 2, \dots, l \quad (6)$$

No qual l é o número total de neurônios na grade. Para encontrar o melhor casamento do vetor de entrada x com os vetores de pesos sinápticos w_j , compare os produtos internos $w_j^T x$ para $j=1, 2, \dots, l$ e selecione o maior. Isto assume que o mesmo limiar seja aplicado a todos os neurônios; o limiar é o negativo das *bias*. Assim, selecionando o neurônio com o maior produto interno $w_j^T x$, teremos de fato determinado a localização onde a vizinhança topológica dos neurônios excitados deve ser centrada (HAYKIN, 2001; KOHONEN, 2013).

O melhor casamento, baseado na maximização do produto interno $w_j^T x$, é matematicamente equivalente a minimizar a distância euclidiana entre os vetores x e w_j . Se for usado o índice $i(x)$ para identificar o neurônio que melhor casa com o vetor de entrada x , pode-se então determinar $i(x)$ aplicando a condição mostrada na Equação (7), que resume a essência do processo competitivo entre os neurônios (KOHONEN, 2013; DEMIR e CERGIBOZAN, 2018).

$$i(x) = \arg \min_j \|x - w_j\|, j = 1, 2, \dots, l \quad (7)$$

Dependendo da aplicação de interesse, a resposta da grade pode ser tanto o índice do neurônio vencedor, como o vetor de peso sináptico que está mais próximo do vetor de entrada em um sentido euclidiano (HAYKIN, 2001; KOHONEN, 2001; KOHONEN, 2013).

A Figura 10 ilustra um mapa auto organizável. Com a entrada de dados X transmitida para um conjunto de modelos M_i dos quais M_c combina melhor com X . Todos os modelos que ficam na vizinhança (círculo maior) de M_c na grade combinam melhor com X do que com o restante.

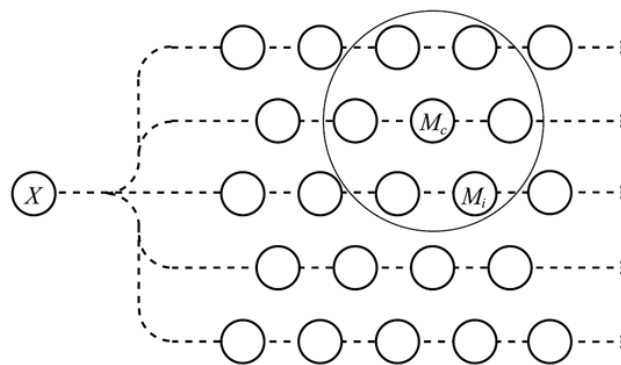


Figura 10 – Modelo de vizinhança da rede SOM. Fonte: Kohonen (2013).

O critério para escolha do neurônio vencedor pode-se basear em qualquer métrica que informe a distância entre o vetor de pesos dos neurônios. A medida de distância mais utilizada é a distância euclidiana, porém outras medidas podem ser utilizadas.

O cálculo da distância euclidiana, que é a distância entre dois pontos medida em linha reta que liga esses dois pontos utiliza a Equação (8) (DEZA e DEZA, 2016).

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (8)$$

Outra distância utilizada é a manhattan, que é a distância entre dois pontos medida ao longo dos eixos coordenados em ângulos retos utiliza a Equação (9) (DEZA e DEZA, 2016).

$$d(x, y) = \sum_{k=1}^n |x_k - y_k| \quad (9)$$

Se os vetores de entrada e de pesos já estiverem normalizados de forma unitária, o processo de aprendizagem obtém maior eficiência. A normalização unitária é efetuada

bastando-se dividir cada vetor pelo seu respectivo módulo (SILVA, SPATTI e FLAUZINO 2010).

A Figura 11 mostra a regra de ajuste de pesos para o neurônio vencedor do processo competitivo, em que três neurônios representados por: $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ que disputam a competição frente à apresentação do padrão $x^{(k)}$, sendo o mesmo composto por duas entradas x_1 e x_2 .

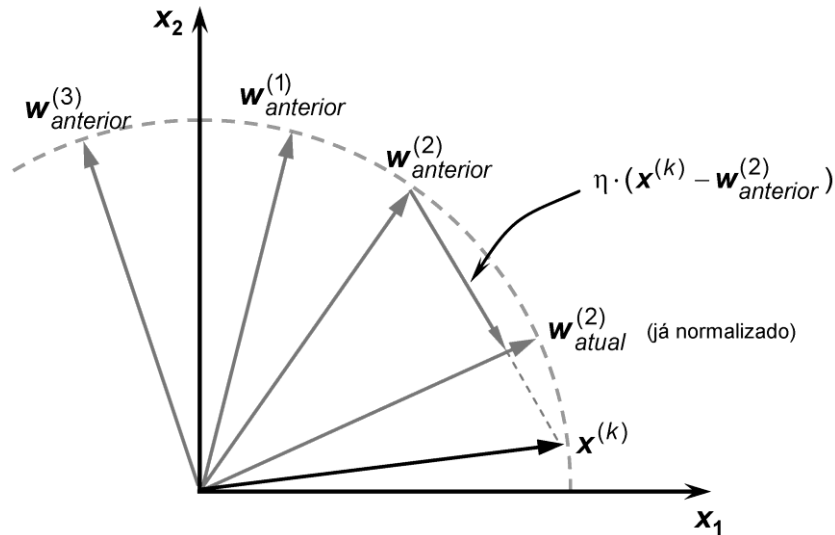


Figura 11 – Processo de ajuste do vetor de pesos do neurônio vencedor. Fonte: Silva, Spatti e Flauzino (2010).

A partir da Figura 11 observa-se que todos os vetores se encontram normalizados, o vetor $w^{(2)}$, é o vencedor da competição com os vetores $w^{(1)}$ e $w^{(3)}$, pois é o que está mais próximo do padrão $x^{(k)}$. Verifica-se que o ajuste consistiu tão somente em rotacionar o neurônio vencedor em direção ao vetor representando a amostra (SILVA, SPATTI e FLAUZINO 2010).

Conforme Demir e Cergibozan (2018) o neurônio vencedor é aquele que tiver o maior nível de ativação u_i , como o nível de ativação de um neurônio i é o produto entre o vetor de entrada $x^{(k)}$ e o vetor de pesos do neurônio $w^{(i)}$, conforme a Equação (10).

$$u_i = w^{(i)} x^{(k)} \quad (10)$$

O critério para a escolha do neurônio vencedor é o de similaridade entre $x^{(k)}$ e $w^{(i)}$. Este produto pode ser escrito conforme a Equação (11).

$$u_i = |w^{(i)}| |x^{(k)}| \cos \theta \quad (11)$$

onde θ é o ângulo entre os vetores $x^{(k)}$ e $w^{(i)}$.

A Figura 12 ilustra quatro classes disponíveis, em que se observa a distribuição dos vetores de pesos, após a estabilização da rede SOM. Cada um destes vetores são representativos das amostras do problema, é constituído por dois sinais de entrada x_1 e x_2 .

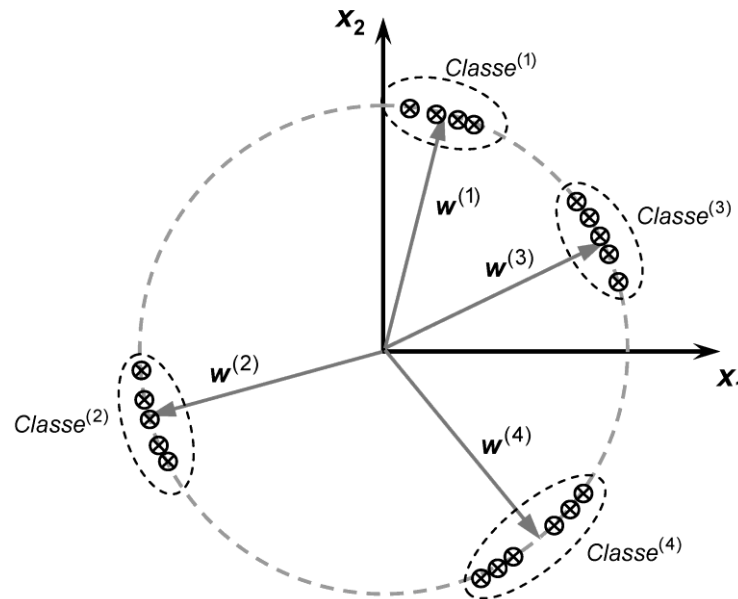


Figura 12 – Distribuição em espaço bidimensional dos vetores. Fonte: Silva, Spatti e Flauzino (2010).

Percebe-se na Figura 12, assim como na Figura 11, que todos os vetores acabam localizados no círculo unitário, à medida que são bidimensionais e estão unitariamente normalizados. Percebe-se, ainda, que os quatro vetores de pesos se posicionam nos centros dos agrupamentos que representam as amostras (\otimes).

Desta forma, toma-se conhecimento que a estrutura neural representada nesta figura foi constituída por quatro neurônios, pois há somente quatro vetores de pesos $w^{(1)}$, $w^{(2)}$, $w^{(3)}$, $w^{(4)}$. (HAYKIN, 2001).

Constata-se ainda na Figura 12 que, se a estrutura neural fosse constituída por cinco neurônios, ao invés de quatro, as amostras da classe⁽⁴⁾ seriam divididas em duas classes, pois a mesma classe está agrupando dois subconjuntos, embora muito próximos, aparecem em apenas um agrupamento.

2.3.5 O Processo Cooperativo da rede SOM

No processo cooperativo o neurônio vencedor localiza o centro de uma vizinhança topológica de neurônios cooperativos. Há evidência neurobiológica para a interação lateral

entre um conjunto de neurônios excitados. Em particular um neurônio que está disparando tende a excitar mais forte os neurônios na sua vizinhança imediata que aqueles mais distantes dele o que é intuitivamente razoável. Isso faz com que a vizinhança topológica em torno do neurônio vencedor decaia suavemente com a distância lateral (HAYKIN, 2001).

Observa-se que $h_{j,i}$ representa a vizinhança topológica centrada no neurônio vencedor i e que contenha um conjunto de neurônios excitados ou cooperativos, sendo um neurônio típico deste conjunto representado por j . Considera-se que $d_{i,j}$ represente a distância lateral entre o neurônio vencedor i e o neurônio excitado j . Então pode-se assumir que a vizinhança topológica $h_{j,i}$ é uma função unimodal da distância $d_{j,i}$, desde que ela satisfaça duas exigências distintas:

- A vizinhança topológica $h_{j,i}$ é simétrica em relação ao ponto máximo definido por $d_{j,i} = 0$; em outras palavras, ela alcança o seu valor máximo no neurônio vencedor i para o qual a distância de $d_{j,i}$ é zero.
- A amplitude da vizinhança topológica $h_{j,i}$ decresce monotonamente com o aumento da distância lateral $d_{j,i}$, decaindo a zero para $d_{j,i} \longrightarrow \infty$; esta é condição necessária para a convergência (DEMIR e CERGIBOZAN, 2018).

A função que atende a estas exigências é a função gaussiana apresentada na Equação (12), sendo comum o seu uso.

$$h_{j,i(x)} = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right) \quad (12)$$

A Figura 13 ilustra o neurônio central rotulado v que é o neurônio vencedor. O parâmetro σ é a largura efetiva da vizinhança topológica.

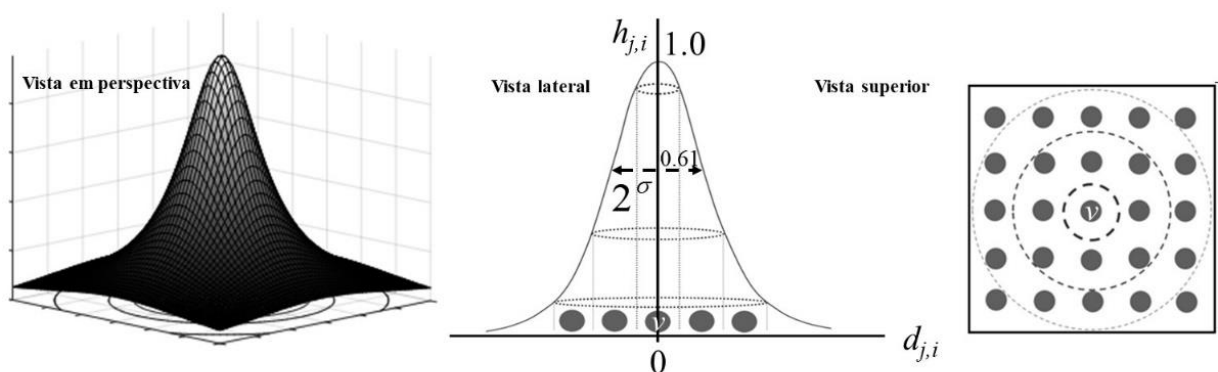


Figura 13 – Função de vizinhança gaussiana apresentada sob diferentes vistas. Fonte: Adaptado de Haykin (2001); Silva, Peres e Boscaroli (2016).

A função de vizinhança topológica gaussiana faz com que o algoritmo da rede SOM convirja com maior rapidez. No sentido qualitativo, a vizinhança topológica hexagonal oferece tradicionalmente resultados melhores que a vizinhança retangular (HAYKIN, 2001).

A função de vizinhança topológica mede o grau com o qual os neurônios excitados na vizinhança do neurônio vencedor participam do processo de aprendizagem (DEMIR e CERGIBOZAN, 2018).

A Figura 14 mostra um exemplo de vizinhança topológica retangular sendo ($t_1 < t_2 < t_3$). No centro da vizinhança topológica observa-se o neurônio vencedor.

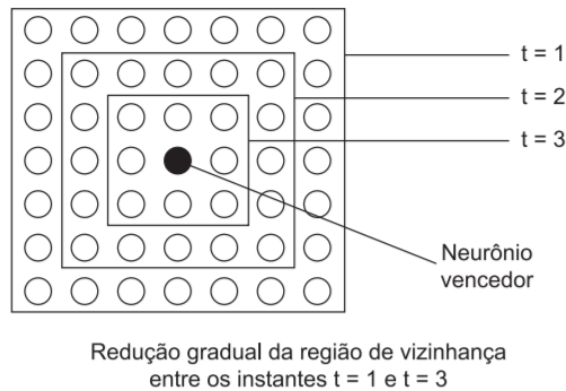


Figura 14 – Vizinhança topológica retangular. Fonte: Adaptado de Kohonen (2001).

Para que a cooperação entre os neurônios vizinhos se mantenha, é necessário que a vizinhança topológica $h_{j,i}$ seja dependente da distância lateral $d_{j,i}$ entre o neurônio vencedor i e o neurônio excitado j no espaço de saída em vez de ser dependente de alguma medida de distância no espaço de entrada original, conforme verifica-se na Equação (12).

Na grade unidimensional, $d_{j,i}$ é um inteiro igual a $|j-i|$, e na grade bidimensional é definida pela Equação (13),

$$d_{j,i}^2 = \|\mathbf{r}_j - \mathbf{r}_i\|^2 \quad (13)$$

em que o vetor discreto \mathbf{r}_j define a posição do neurônio excitado j e \mathbf{r}_i define a posição discreta do neurônio vencedor i , sendo ambos medidos no espaço de saída discreto (HAYKIN, 2001).

Outra característica única do algoritmo SOM é que o tamanho da vizinhança topológica diminui com o tempo. Esta exigência é satisfeita fazendo-se com que a largura σ da função de vizinhança topológica $h_{j,i}$ diminua com o tempo.

Uma escolha típica para a dependência de σ com o tempo discreto n é o decaimento exponencial conforme mostra a Equação (14),

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{\tau_1}\right), n = 0, 1, 2, \dots, \quad (14)$$

Sendo que, σ_0 é o valor de σ na inicialização do algoritmo SOM, e τ_1 é a uma constante de tempo. Consequentemente, a vizinhança topológica assume uma forma variável no tempo, como mostra a Equação (15).

$$h_{j,i(x)}(n) = \exp\left(\frac{d_{j,i}^2}{2\sigma^2(n)}\right), n = 0, 1, 2, \dots, \quad (15)$$

Sendo que, $\sigma(n)$ é definido pela Equação (14).

Assim, quando o número de iterações n aumenta, a largura $\sigma(n)$ decresce a uma taxa exponencial e a vizinhança topológica diminui de maneira correspondente.

Outra forma de verificar a variação da função de vizinhança $h_{j,i(x)}(n)$ em torno de um neurônio vencedor $i(x)$ é o propósito de um $h_{j,i(x)}(n)$ largo, é essencialmente correlacionar as direções das atualizações dos pesos de um grande número de neurônios excitados da grade. Quando a largura de $h_{j,i(x)}(n)$ é diminuída, também diminui o número de neurônios cujas direções de atualização são correlacionadas.

Segundo Haykin (2001), é um desperdício de recursos computacionais mover um grande número de graus de liberdade em torno de um neurônio vencedor de forma correlacionada, como no caso do algoritmo SOM padrão.

É melhor utilizar uma forma de treinamento SOM normalizada, na qual usa-se um número muito menor de graus de liberdade normalizados, esta operação é facilmente realizada na forma discreta tendo uma função de vizinhança $h_{j,i(x)}(n)$ de largura constante, mas aumentando gradualmente o número total de neurônios (HAYKIN, 2001).

Os novos neurônios são inseridos na metade da distância entre os neurônios antigos, e a suavidade do algoritmo SOM garante que os novos neurônios se insiram na adaptação sináptica de maneira mais suave.

2.3.6 O Processo Adaptativo Sináptico da rede SOM

O último processo envolvido na formação do mapa auto organizável é o processo adaptativo sináptico.

Para que a grade seja auto organizável, é necessário que o vetor de peso sináptico w_j do neurônio j da grade se modifique em relação ao vetor de entrada x .

No postulado de aprendizagem de Hebb (1949), um peso sináptico é aumentado com uma ocorrência simultânea de atividades pré-sináptica e pós-sináptica, o uso de tal regra é muito adequado para aprendizagem associativa. Entretanto, para o tipo de aprendizagem não-supervisionada, considerado a hipótese hebbiana na sua forma básica, não é satisfatória pelas seguintes razões: as modificações das conectividades ocorrem apenas em uma direção, o que leva no final todos os pesos à saturação (HAYKIN, 2001).

Para superar este problema, modifica-se a hipótese hebbiana, incluindo um termo de esquecimento – $g(y_j)w_j$ onde w_j é o vetor peso sináptico do neurônio j e $g(y_j)$ é uma função escalar positiva da resposta y_j . A única exigência imposta à função $g(y_j)$ é que o termo constante da expansão de $g(y_j)$ seja zero, de modo que se pode escrever de acordo a Equação (16).

$$g(y_j) = 0 \quad \text{para} \quad y_j = 0 \quad (16)$$

Segundo Haykin (2001), o significado desta exigência será explicado a seguir. Dada esta função, pode-se então expressar a modificação do vetor peso do neurônio j da grade como mostra a Equação (17),

$$\Delta w_j = \eta y_j x - g(y) w_j \quad (17)$$

onde η é o parâmetro da taxa de aprendizagem do algoritmo, o primeiro termo do lado direito da Equação (17) é o termo hebbiano e o segundo termo é o termo de esquecimento.

Para atender a exigência da Equação (16), foi utilizada uma função linear para $g(y_j)$, como mostra a Equação (18).

$$g(y_j) = \eta y_j \quad (18)$$

É possível simplificar mais a Equação (18) conforme a Equação (19):

$$y_j = h_{j,i(x)} \quad (19)$$

Utilizando as Equações (18) e (19) na Equação (17) chega-se a:

$$\Delta w_j = \eta^h_{j,i(x)} (x - w_j) \quad (20)$$

Por fim, utilizando o formalismo de tempo discreto, dado o vetor de peso sináptico $w_j(n)$ do neurônio j no tempo n , o vetor de peso atualizado $w_j(n+1)$ no tempo $n+1$ é definido por Kohonen (2001), conforme a Equação (21),

$$w_j(n+1) = w_j(n) + \eta(n) h_{j,i(x)}(n) (x - w_j(n)) \quad (21)$$

A Equação (21) é aplicada a todos os neurônios da grade que se encontram dentro da vizinhança topológica do neurônio vencedor i , e tem o efeito de mover o vetor peso sináptico w_i do neurônio i em direção ao vetor de entrada x .

Para Haykin (2001), o parâmetro da taxa de aprendizagem $\eta(n)$ deve ser variável no tempo como mostrado na Equação (21) anteriormente, que corresponde ao caso da aproximação aleatória.

A taxa de aprendizagem possui valores entre 0 e 1 (CAMPOS e SAITO, 2004) e deve começar com um valor inicial η_0 e então descrever gradualmente com o aumento do tempo n , esta exigência pode ser satisfeita escolhendo-se um decaimento exponencial para $\eta(n)$, como mostra a Equação (22).

$$\eta(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right), n = 0,1,2,\dots, \quad (22)$$

onde τ_2 é uma outra constante de tempo do algoritmo SOM. Apesar das fórmulas de decaimento exponencial descritas nas Equações (14) e (22) para largura da função de vizinhança e o parâmetro da taxa de aprendizagem, respectivamente, poderem não ser ótimas, elas são normalmente adequadas para a formação do mapa de características de uma maneira auto organizada (HAYKIN, 2001).

Num estado inicial de desordem completa, é surpreendente como o algoritmo SOM gradualmente leva a uma representação organizada de padrões de ativação retirados do espaço

de entrada, desde que os parâmetros do algoritmo sejam selecionados adequadamente (HAYKIN, 2001).

De acordo com Kohonen (2001), é possível decompor a adaptação dos pesos sinápticos da grade, calculada de acordo com a Equação 4.16, em duas fases: uma fase de ordenação ou de auto-organização seguida por uma fase de convergência, estas duas fases do processo adaptativo da rede SOM são descritas a seguir:

- a) Fase de auto-organização ou de ordenação. É durante esta fase do processo adaptativo que ocorre a ordenação topológica dos vetores de peso.

Segundo Haykin (2001) a fase de ordenação pode exigir mil iterações do algoritmo SOM, em geral mais. Nesta fase deve-se ter cautela na escolha do parâmetro de aprendizagem e da função de vizinhança.

- b) Fase de convergência. Esta fase do processo adaptativo é necessária para realizar uma sintonia fina do mapa de características e assim produzir uma quantização estatística precisa do espaço de entrada.

O número de iterações que constituem a fase de convergência deve ser no mínimo quinhentas vezes o número de neurônios na grade. Assim, a fase de convergência pode durar milhares ou dezenas de milhares de iterações.

2.3.7 Medidas de Qualidade da rede SOM

a) Erro de Quantização (EQ) corresponde à média das distâncias entre cada vetor de dados (x) e o correspondente vetor de pesos (w) do neurônio BMU (*Best Match Unit*). A medida corresponde à resolução do mapa, valores menores indicam um ajuste melhor. Kohonen (2001) sugeriu o EQ como a medida básica de qualidade para avaliar a auto-organização do mapa. Para calcular o EQ utiliza-se a Equação (23) (KIVILUOTO, 1996; BREARD, 2017).

$$EQ = \frac{1}{N} \sum_{k=1}^N \|x - w_i\| \quad (23)$$

Estima-se que quanto menor o EQ, mais bem ajustado o neurônio vencedor estará aos vetores de entrada.

b) Erro Topográfico (ET). Um dos principais objetivos da rede SOM é preservar os recursos topológicos do espaço de entrada no espaço de saída de baixa dimensionalidade. O ET quantifica a capacidade do mapa em representar a topologia dos dados de entrada é uma medida

de quão bem a estrutura do espaço de entrada foi modelada pelo mapa. Especificamente, avalia as descontinuidades locais no mapeamento (KIVILUOTO, 1996; BREARD, 2017).

Para cada vetor de dados de entrada (x) são calculados seu primeiro BMU e o seu segundo BMU e toda vez que eles não forem vizinhos, aumenta-se o erro em uma unidade, tirando-se depois a média pelo número total de vetores, onde $u(x_k)$ igual a 1 se o primeiro e o segundo BMU não são vizinhos, caso contrário igual a 0 (KIVILUOTO, 1996; BREARD, 2017).

A Equação (24) calcula o erro topográfico.

$$ET = \frac{1}{N} \sum_{k=1}^N u(x_k) \quad (24)$$

Quando os valores do EQ e ET são muito baixos, suspeita-se de *overfitting*. A rede SOM, na tentativa de representar o mais fielmente possível os dados, dobra-se de forma que acaba representando exatamente os dados, podendo perder a sua capacidade de generalização (KASKI e KOHONEN, 1997).

O fenômeno inverso, o *underfitting*, ocorre quando um mapa é rígido demais. Isso pode acontecer quando há poucos neurônios para representar um número proporcionalmente grande de dados ou o raio de vizinhança final da função gaussiana for maior que 1 durante o treinamento. Neste caso, os valores de EQ podem ser mais altos, os vetores de pesos dos neurônios encontram-se, em média, menos próximos dos vetores de dados (KASKI e KOHONEN, 1997).

A rede SOM possui algumas vantagens em relação aos demais métodos de agrupamento, pois fornece uma visualização bidimensional dos agrupamentos independentemente da dimensão da base de dados, robusto na presença de dados com ruídos e *outliers* e não precisam previamente do número de agrupamentos. O mapa auto organizável gerado pela rede SOM é uma ferramenta excelente na fase exploratória da mineração de dados, tendo em vista que projeta o espaço de entrada multidimensional em um mapa bidimensional o que pode ser efetivamente usado para visualizar e explorar propriedades dos dados (VESANTO e ALHONIEMI, 2000).

2.3.8 Teoria dos Conjuntos Aproximados

Rough Sets ou Conjuntos aproximados podem ser considerados conjuntos com

fronteiras nebulosas, ou seja, conjuntos que não podem ser caracterizados precisamente, utilizando-se dos atributos disponíveis (PAWLAK, 1982).

RS trata de um tipo fundamental de incerteza, a indiscernibilidade. A indiscernibilidade surge quando não é possível distinguir elementos de um mesmo conjunto, e representa a situação em que esses elementos parecem todos serem um único elemento (UCHÔA, 1998).

Os conceitos de RS têm se mostrado muito úteis quando aplicados a problemas do tipo: redução de atributos, descoberta de dependência entre atributos e na descoberta de padrões entre os dados (PAWLAK, 1982).

A redução de atributos realizada pelos RS é feita através dos chamados redutos, que são subconjuntos de atributos capazes de representar o conhecimento da base de dados com todos os seus atributos iniciais. Este procedimento de eliminação de atributos irrelevantes é uma das características da Teoria.

2.3.8.1 Principais Conceitos

Este trabalho não tem como finalidade o aprofundamento no formalismo matemático dos RS, que é grande e considerado um aspecto importante da Teoria dos *Rough Sets*. Para isto recomenda-se o trabalho de Uchôa (1998). Este formalismo matemático ainda não é sistematicamente abordado, não é padronizado e tampouco muito explorado. Ainda não existe uma padronização da notação matemática. Assim, foi necessário adotar a notação matemática proposta por Sassi (2006), com o objetivo de torná-la mais clara.

Os principais conceitos dos RS com a notação matemática pertinente estão relacionados seguir: Aproximação Inferior (AI) \underline{B} , Aproximação Superior \bar{B} (AS), Sistema de Informação (S), Atributos Condicionais (C), Atributo de Decisão (d), Sistema de Decisão (SD), Região de Fronteira $RF(X)$, Região Negativa $RN(X)$, Reduto (RED) e Indiscernibilidade (IND). Adotou-se que a palavra elementos é tratada como sinônimo de casos, exemplos ou registros que compõem uma base de dados.

- Espaços Aproximados

Um espaço aproximado é um par ordenado $A = (U, R)$, onde: U é um conjunto não vazio, denominado conjunto universo, e R é uma relação de equivalência sobre U , denominada Relação de Indiscernibilidade. Uma relação binária $R \subseteq X \times X$, a qual é reflexiva (um elemento está relacionado com ele próprio $x R x$), simétrica (se $x R y$ então $y R x$) e transitiva (se $x R y$ e $y R z$ então $x R z$), é chamada de relação de equivalência.

Dados os elementos $x, y \in U$, se $x R y$ então x e y são indiscerníveis em A , ou seja, a

classe de equivalência definida por x é a mesma que a definida por y , i.e., $[x]R = [y]R$. A classe de equivalência de um elemento $x \in X$ consiste em todos os elementos $y \in X$ para os quais xRy . Os elementos que são indiscerníveis formam conjuntos chamados de conjuntos elementares. Dessa forma, pode-se dizer que as classes de equivalência de R são os conjuntos elementares de A . Na Figura 15 pode-se visualizar o espaço aproximado $A = (U, R)$.

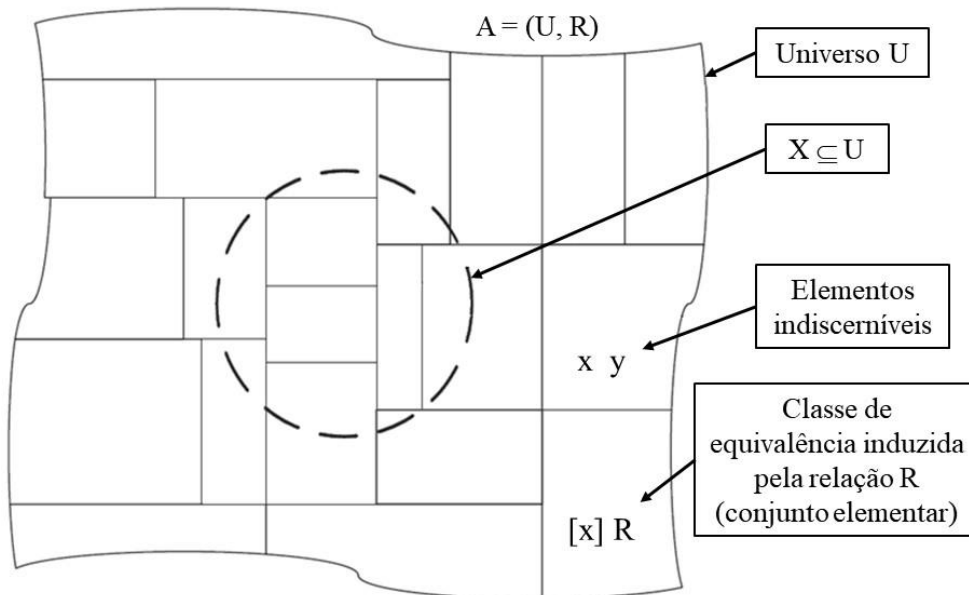


Figura 15 – Espaço Aproximado $A = (U, R)$ e $X \subseteq U$. Fonte: Adaptada de Sassi (2006); Pawlak e Skowron (2007).

A forma mais comum para representação dos dados em RS é através de um sistema de informação (S) que contém um conjunto de elementos, sendo que cada elemento tem uma quantidade de atributos condicionais. Esses atributos são os mesmos para cada um dos elementos, mas os seus valores nominais podem diferir (Tabela 1).

Dessa forma, um sistema de informação é um par ordenado $S = (U, C)$, onde U é um conjunto finito e não-vazio de elementos chamado de universo (Figura 15), e C é um conjunto finito e não-vazio formado pelos atributos. Cada atributo $a \in C$ é uma função $a: U \rightarrow V_a$, onde V_a é o conjunto dos valores permitidos para o atributo a (sua faixa de valores).

Na Tabela 1, onde é apresentado o sistema de informação S , podem-se observar os principais conceitos de RS, o espaço aproximado $A = (U, R)$, o universo U formado pelos elementos Ação1; Ação2; Ação3; Ação4; Ação5; Ação6 e os Atributos (C) Fechamento, Nova Mínima, Preço Máximo e R a relação de equivalência sobre U .

Tabela 1 – Exemplo de um Sistema de Informação (S)

Elementos - Ações	Preço de Fechamento	Nova Mínima	Preço Máximo
Ação1 {e1}	Positivo	Não	Não Rompeu
Ação2 {e2}	Neutro	Não	Não Rompeu
Ação3 {e3}	Neutro	Não	Não Rompeu
Ação4 {e4}	Negativo	Sim	Não Rompeu
Ação5 {e5}	Neutro	Sim	Rompeu
Ação6 {e6}	Positivo	Não	Rompeu

Fonte: Adaptado de Sassi (2006).

- Indiscernibilidade

O principal conceito envolvido em RS é a relação de indiscernibilidade (PAWLAK, 1982) a qual normalmente está associada a um conjunto de atributos. Se tal relação existe entre dois elementos, isso significa que todos os valores nominais dos seus atributos são idênticos com respeito aos atributos considerados, portanto não podem ser discernidos (distinguidos) entre si.

Para cada subconjunto de atributos $B \subseteq C$ no sistema de informação $S = (U, C)$, uma relação de equivalência $INDs(B)$ é associada, chamada de Relação de Indiscernibilidade definida conforme Equação (25):

$$INDs(B) = \{(x,y) \in U^2 / \forall \mathbf{a} \in B, \mathbf{a}(x) = \mathbf{a}(y)\} \quad (25)$$

O conjunto de todas as classes de equivalência na relação $INDs(B)$ é representado por $U/INDs(B)$, denominado quociente de U pela relação $INDs(B)$.

Em muitos casos é importante a classificação dos elementos considerando um atributo de decisão que informa a decisão a ser tomada. Dessa forma, um SI que apresenta um atributo de decisão é denominado Sistema de Decisão (SD).

Um SD pode ser representado por $SD = (U, C \cup \{d\})$, em que $d \notin C$ é o atributo de decisão. A Tabela 2 mostra um SD obtido a partir do sistema de informação S da Tabela 1, destacando os atributos condicionais (Fechamento, Nova Mínima e Preço Máximo) e o atributo de decisão (Tendência).

Para o exemplo ter caráter didático utilizou-se atributos relacionados ao valor (preço) da ação durante o Pregão:

- Preço de Fechamento: que pode assumir dois valores que indicam se o preço da ação

no fechamento do Pregão foi Positivo ou Negativo;

- Nova Mínima: este atributo pode assumir dois valores: Sim ou Não, indicando desta forma se houve um novo valor mínimo, ou seja, se o menor valor atingido durante o Pregão foi menor quando comparado ao Pregão do dia anterior;

- Preço Máximo, este atributo também pode receber dois valores Não Rompeu ou Rompeu, em outras palavras, indica se o maior preço da ação durante o Pregão foi maior (rompeu) ou menor (não rompeu) ao preço da ação no Pregão do dia anterior.

Tabela 2 – Sistema de Decisão (Sistema de Informação com o **Atributo de Decisão:** Tendência)

Elementos - Ações	Atributos Condicionais			Atributo de Decisão
	Preço de Fechamento	Nova Mínima	Preço Máximo	Tendência
Ação1 {e1}	Positivo	Não	Não rompeu	Alta
Ação2 {e2}	Neutro	Não	Não rompeu	Baixa
Ação3 {e3}	Neutro	Não	Não rompeu	Alta
Ação4 {e4}	Negativo	Sim	Não rompeu	Baixa
Ação5 {e5}	Neutro	Sim	Rompeu	Baixa
Ação6 {e6}	Positivo	Sim	Rompeu	Alta

Fonte: Adaptado de Sassi (2006).

Os valores dos atributos são chamados de valores nominais e estão expressos como: Fechamento {Positivo, Neutro, Negativo}; Nova Mínima {Sim, Não}; Preço Máximo {Não Rompeu, Rompeu} e Tendência {Alta, Baixa}. Considerando cada atributo condicional de forma independente, a relação de equivalência do sistema de informação S (Tabela 1) forma os seguintes conjuntos elementares: Fechamento Positivo {Ação1, Ação6}; Neutro {Ação2, Ação 3, Ação5}; Negativo {Ação4}; Nova Mínima: Não {Ação1, Ação2, Ação3}; Sim {Ação4, Ação5, Ação6} e Preço Máximo: Não Rompeu {Ação1, Ação2, Ação3, Ação4}; Rompeu {Ação5, Ação6}.

Ao utilizar todos os atributos condicionais do sistema de informação S da Tabela 1 obtêm-se os seguintes conjuntos elementares: {Ação1}, {Ação2, Ação3}, {Ação4}, {Ação5} e {Ação6}. Na Tabela 3, pode-se perceber que existem 2 (dois) elementos (casos) {Ação2} e {Ação3} iguais (destacados em negrito), no que se refere à valores de atributos condicionais.

Tabela 3 – Sistema de Decisão com os elementos Ação2 e Ação3 indiscerníveis, com relação aos atributos condicionais

Elementos - Ações	Preço de Fechamento	Nova Mínima	Preço Máximo	Tendência
Ação1 {e1}	Positivo	Não	Não rompeu	Alta
Ação2 {e2}	Neutro	Não	Não rompeu	Baixa
Ação3 {e3}	Neutro	Não	Não rompeu	Alta
Ação4 {e4}	Negativo	Sim	Não rompeu	Baixa
Ação5 {e5}	Neutro	Sim	Rompeu	Baixa
Ação6 {e6}	Positivo	Sim	Rompeu	Alta

Fonte: Adaptado de Sassi (2006).

Existindo a Relação de Indiscernibilidade entre os elementos {Ação2} e {Ação3} como mostrado na Tabela 3, significa que todos os valores nominais de seus atributos são idênticos com relação ao subconjunto de atributos B ($B \subseteq S$) considerado, ou seja, são indiscerníveis, não podem ser diferenciados entre si.

- Aproximação de Conjuntos

A Tabela 3 apresenta os elementos do sistema de informação S segundo as características do atributo de decisão. Pode-se então, fazer a seguinte pergunta: Quais características dos atributos condicionais definem a tendência de baixa ou de alta? Nota-se que não há uma resposta única para esta pergunta, pois as ações {2} e {3} apresentam as mesmas características dos atributos condicionais, mas se diferenciam no atributo de decisão.

Pode-se dizer com certeza, conforme a Tabela 3, que qualquer ação com características iguais as das ações {1} e {6} terão lucro, assim como qualquer ação com características iguais as das ações {4} ou {5} terá prejuízo, porém, nada se pode afirmar para ações com características iguais as das ações {2} e {3}, pois, apesar de apresentarem atributos condicionais com as mesmas características, possuem atributos de decisão diferentes. São nesses casos que RS pode ser aplicado.

Um conjunto definível em S é qualquer união finita de conjuntos elementares. Para cada conceito X que é o conjunto de elementos com respeito a B , ou seja, X é obtido através das informações dos atributos de B , são computados o maior conjunto definível contido em X e o menor conjunto definível que contém X . O primeiro conjunto é chamado de Aproximação Inferior (AI) de X , enquanto o segundo conjunto é chamado de Aproximação Superior (AS) de X (PAWLAK, 1982).

A Aproximação Inferior $\underline{B}(X)$ e a Aproximação Superior $\overline{B}(X)$ de um conjunto de elementos $X \subseteq U$ com respeito a um conjunto de atributos $B \subseteq S$ (definindo uma relação de

equivalência em U) podem ser definidas em termos das classes na relação de equivalência, conforme as Equações (26) e (27).

$$\underline{B}(X) = \{x \in U \mid U/\text{INDs}(B) \subseteq X\} \quad (26)$$

$$\bar{B}(X) = \{x \in U \mid U/\text{INDs}(B) \cap X \neq \emptyset\} \quad (27)$$

Os elementos da Aproximação Inferior $\underline{B}(X)$ são classificados com certeza como membros de X , utilizando o conjunto de atributos B , enquanto os elementos da Aproximação Superior $\bar{B}(X)$ podem ser classificados como possíveis membros de X , utilizando o mesmo conjunto de B . Os elementos que correspondem a alta podem ser tomados como exemplos do conceito X . Com base na Tabela 4 pode-se observar que existem três elementos que possuem como atributo de decisão alta $\{e1, e3, e6\}$, porém, existe Relação de Indiscernibilidade entre os elementos $\{e2\}$ e $\{e3\}$, impedindo que $\{e3\}$ seja considerado com certeza como membro de X . Assim, somente os elementos $\{e1, e6\}$ podem ser classificados como membros de X e elementos da Aproximação Inferior $\underline{B}(X)$. A Figura 16 ilustra a Aproximação Inferior $\underline{B}(X)$ destacada na cor cinza.

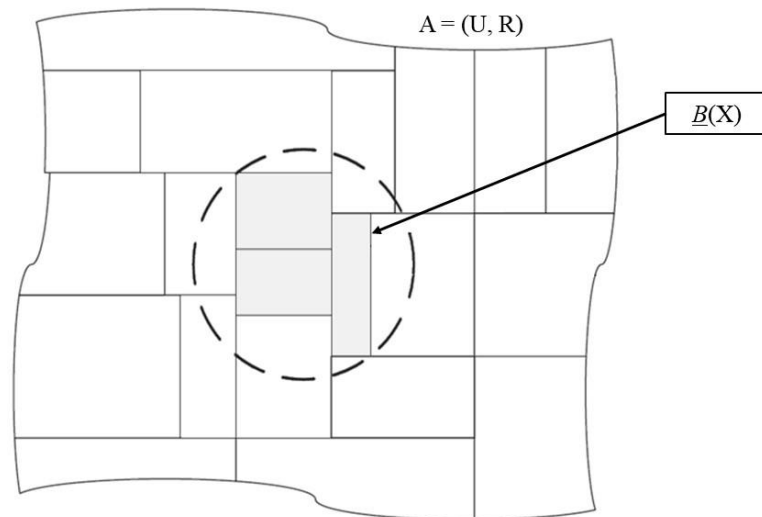


Figura 16 – Aproximação Inferior de X . Fonte: Adaptada de Sassi (2006).

Na Figura 16, nos quadrados de cor cinza, estão contidos os elementos que correspondem a Aproximação Inferior $\underline{B}(X)$, os quadrados em branco tocados pela elipse (X) estão os elementos que correspondem a Aproximação Superior $\bar{B}(X)$ e os quadrados em branco que não são tocados pela elipse correspondem com certeza aos elementos que não pertencem a $\bar{B}(X)$ (Região Negativa).

Como já foi dito, na Aproximação Superior $\bar{B}(X)$ são classificados os elementos que são possíveis membros de X . Dessa forma, a Aproximação Superior reúne os elementos com

atributo de decisão igual à alta {e1, e3, e6} e também os elementos que possuem atributo de decisão igual à baixa, desde que exista uma Relação de Indiscernibilidade entre os elementos, como ocorre entre o elemento {e2} e o elemento {e3}. A Figura 17 ilustra a Aproximação Superior (X).

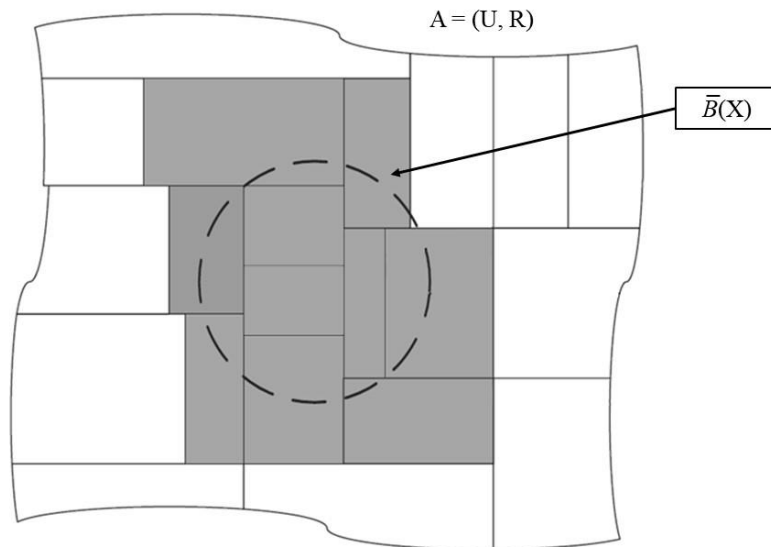


Figura 17 – Aproximação Superior de X. Fonte: Adaptada de Sassi (2006).

Portanto, obtém-se uma Região de Fronteira de X pela diferença de $\bar{B}(X)$ para $\underline{B}(X)$, representada por $RF(X)$, ou seja, $RF(X) = \bar{B}(X) - \underline{B}(X)$. A Região de Fronteira, também chamada de Duvidosa, possui somente os elementos que não podem ser classificados com certeza como pertencentes em X, utilizando o conjunto de atributos B. É a região formada pelos elementos de U que pertencem à Aproximação Superior, mas que não pertencem à Aproximação Inferior.

Um conjunto X é definido como *rough* (impreciso) se a sua Região de Fronteira é diferente do conjunto vazio ($RF(X) \neq \emptyset$), e é definido como *crisp* (preciso) se o conjunto for vazio ($RF(X) = \emptyset$). Na Tabela 3, pode-se observar que os elementos {e2, e3} fazem parte da Região de Fronteira.

A Região Negativa é dada pela diferença dos elementos de U para $\bar{B}(X)$, representada por $RN(X) = U - \bar{B}(X)$. A Região Negativa possui somente os elementos que com certeza não podem ser classificados como pertencentes à Aproximação Superior $\bar{B}(X)$, utilizando o conjunto de atributos B.

As regiões de X ficaram assim:

- Aproximação Inferior: $\underline{B}(X)$ {e1, e6}

- Aproximação Superior: $\bar{B}(X)$ {e1, e2, e3, e6}
- Região de Fronteira (Duvidosa): $RF(X)$ {e2, e3}
- Região Negativa: $RN(X)$ {e4, e5}

A Figura 18 ilustra todas as regiões de X em A.

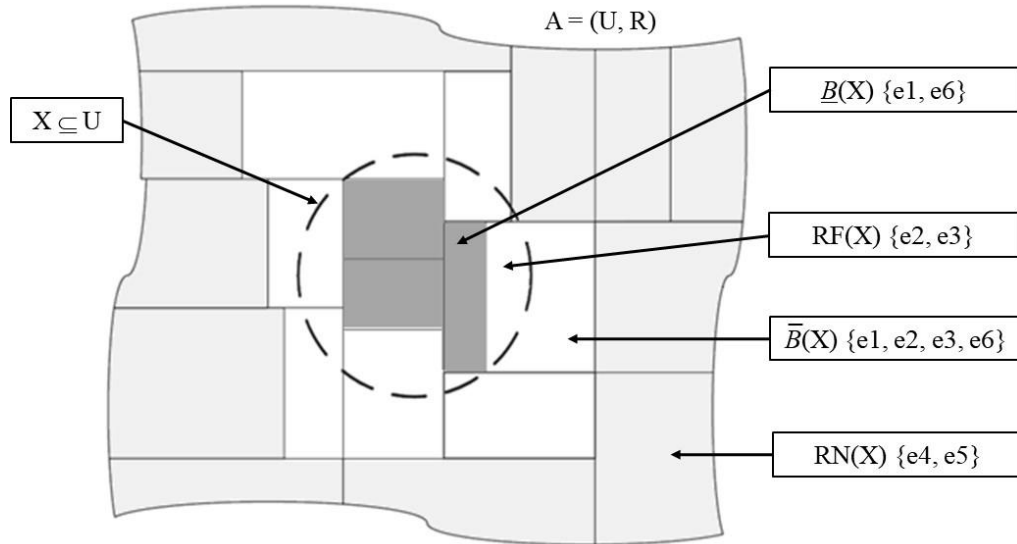


Figura 18 – Todas as regiões de X em A. Fonte: Adaptada de Sassi (2006).

Por meio dos conceitos de AS e AI, podem-se definir as quatro classes básicas de incerteza em RS (KOMOROWSKI *et al.*, 1999).

Um conjunto definível em S é qualquer união finita de conjuntos elementares. O significado dessas classes, em que $-X$ denota $U - X$ (complemento) é o seguinte:

X é *rough* B-definível, se e somente se $\underline{B}(X) \neq \emptyset$ e $\bar{B}(X) \neq U$.

Se X é *rough* B-definível, isso significa que é possível decidir para alguns elementos de U quando eles pertencem a X e para alguns elementos de U quando eles pertencem a $-X$, utilizando B:

X é internamente B-indefinível, se e somente se $\underline{B}(X) = \emptyset$ e $\bar{B}(X) \neq U$.

Se X é internamente B-indefinível, isso significa que é possível decidir para alguns elementos de U quando eles pertencem a $-X$, mas não é possível decidir para nenhum elemento de U quando ele pertence a X, utilizando B;

X é externamente B-indefinível, se e somente se $\bar{B}(X) \neq \emptyset$ e $\bar{B}(X) = U$.

Se X é externamente B-indefinível, isso significa que é possível decidir para alguns elementos de U quando eles pertencem a X, mas não é possível decidir para qualquer elemento de U quando ele pertence a $-X$, utilizando B;

X é totalmente B -indefinível, se e somente se $\underline{B}(X) = \emptyset$ e $\overline{B}(X) = U$.

Se X é totalmente B -indefinível, isso significa que não é possível decidir para qualquer elemento de U quando ele pertence a X ou a $\neg X$, utilizando B .

A teoria ainda contém conceitos de Qualidade das Aproximações, Dependência de Atributos e Significância de Atributos, que podem consultados em Sassi (2006).

- Redução de Atributos

A redução de atributos no RS é feita através dos chamados Redutos (RED), que são subconjuntos de atributos capazes de representar o conhecimento da base de dados com todos os seus atributos iniciais (PAWLAK, 1982).

Um Reduto de B sobre um sistema de informação S é um conjunto de atributos $B' \subseteq B$ tal que todos os atributos $a \in (B - B')$ são dispensáveis. Com isso, $U/INDs(B') = U/INDs(B)$. O termo RED(B) é utilizado para denotar a família de redutos de B .

O procedimento para a redução de atributos é o seguinte: comparam-se os conjuntos elementares de cada atributo individualmente e depois dois a dois, três a três, assim sucessivamente com os conjuntos elementares formados com todos os atributos $\{e1\}$; $\{e2, e3\}$; $\{e4\}$; $\{e5\}$; $\{e6\}$. Nesta comparação, quando um atributo apresentar os seus conjuntos elementares iguais aos conjuntos elementares formados com todos os atributos, este atributo pode ser reduzido.

O cálculo de reduções para gerar os redutos é um problema $n-p$ completo, e seu processamento em grandes bases de dados exige grande esforço computacional.

Desta forma, comparando cada atributo individualmente (os conjuntos elementares são formados pelo valor nominal do atributo de decisão) obtêm-se: Preço de Fechamento: Alta $\{e1, e3, e6\}$; Baixa $\{e2, e4, e5\}$; Nova Mínima: Alta $\{e1, e3, e6\}$; Baixa $\{e2, e4, e5\}$; Preço Máximo: Alta $\{e1, e3, e6\}$; Baixa $\{e2, e4, e5\}$.

Nenhum atributo teve os seus conjuntos elementares iguais aos conjuntos elementares formados com todos os atributos, portanto, neste momento não houve redução.

Comparando os atributos: Preço de Fechamento e Nova Mínima = Positivo e Não $\{e1\}$; Neutro e Não $\{e2, e3\}$; Negativo e Sim $\{e4\}$; Neutro e Sim $\{e5\}$; Positivo e Sim $\{e6\}$. O resultado foi: $\{e1\}$; $\{e2, e3\}$; $\{e4\}$; $\{e5\}$; $\{e6\}$.

Os conjuntos elementares obtidos são iguais aos conjuntos elementares obtidos com todos os atributos, portanto, é possível reduzir um atributo.

Comparando os atributos: Preço de Fechamento e Preço Máximo = Positivo e Não Rompeu $\{e1\}$; Neutro e Não Rompeu $\{e2, e3\}$; Negativo e Não Rompeu $\{e4\}$; Neutro e

Rompeu {e5}; Positivo e Rompeu {e6}. O resultado dos conjuntos elementares foi: {e1}; {e2, e3}; {e4}; {e5}; {e6}.

Os conjuntos elementares obtidos são iguais aos conjuntos elementares obtidos com todos os atributos, portanto, existe mais de uma redução de atributo.

Comparando os atributos: Nova Mínima e Preço Máximo = Não e Rompeu {e1, e2, e3}; Sim e Não Rompeu {e4}; Sim e Rompeu {e5, e6}.

O Resultado dos conjuntos elementares foi: {e1, e2, e3}; {e4}; {e5, e6}, mostrando-se diferente dos conjuntos elementares obtidos com todos os atributos e, portanto, não sendo possível nenhuma redução. Chega-se à conclusão de que existem dois redutos RED(B): {Preço de Fechamento e Nova Mínima} e {Preço de Fechamento e Preço Máximo}, ou seja, utilizando qualquer um dos dois redutos consegue-se representar a base de dados com todos os seus atributos originais. Pode-se visualizar na Tabela 4, o sistema de informação S sem a presença do atributo reduzido Preço Máximo.

Tabela 4 – Sistema de Informação S sem o atributo Preço Máximo

Elementos - Ações	Preço de Fechamento	Nova Mínima
Ação1 {e1}	Positivo	Não
Ação2 {e2}	Neutro	Não
Ação3 {e3}	Neutro	Não
Ação4 {e4}	Negativo	Sim
Ação5 {e5}	Neutro	Sim
Ação6 {e6}	Positivo	Não

Fonte: Adaptado de Sassi (2006).

Pode-se visualizar na Tabela 5, o sistema de informação S sem a presença do atributo reduzido Nova Mínima.

Tabela 5 – Sistema de Informação S sem o atributo Nova Mínima

Elementos - Ações	Preço de Fechamento	Preço Máximo
Ação1 {e1}	Positivo	Não rompeu
Ação2 {e2}	Neutro	Não rompeu
Ação3 {e3}	Neutro	Não rompeu
Ação4 {e4}	Negativo	Não rompeu
Ação5 {e5}	Neutro	Rompeu
Ação6 {e6}	Positivo	Rompeu

Fonte: Adaptado de Sassi (2006).

- Matriz de Discernibilidade

A redução de atributos no RS são subconjuntos de atributos capazes de representar o conhecimento da base de dados com todos os seus atributos iniciais (PAWLAK, 1982). Essa redução é feita pela função de discernibilidade, a partir da Matriz de Discernibilidade. Considerando o conjunto de atributos $B = \{\text{Fechamento, Nova Mínima e Preço Máximo}\}$ para o sistema de informação S, o conjunto de todas as classes de equivalência determinadas por B sobre S é dado por $U/INDs(B) = \{\{Ação1\}; \{Ação2, Ação3\}; \{Ação4\}; \{Ação5\}; \{Ação6\}\}$, que estão representadas na Tabela 6.

A Matriz de Discernibilidade do sistema de informação S, denotada por $MD(B)$, é uma matriz simétrica $n \times n$ com: $mD(i, j) = \{a \in B \mid a(E_i) \neq a(E_j)\}$ para $i, j = 1, 2, \dots, n$. sendo $1 \leq i, j \leq n$ e $n = |U / INDs(B)|$. Logo, os elementos da matriz de discernibilidade $mD(i, j)$ é o conjunto de atributos condicionais de B que diferenciam os elementos das classes com relação aos seus valores nominais.

Considerando Fechamento (FT), Nova Mínima (NM) e Preço Máximo (PM), com a finalidade de construir a Matriz de Discernibilidade MD (B), tem-se na Tabela 6 a sua representação.

Tabela 6 – Matriz de Discernibilidade

	Ação1	Ação2	Ação3	Ação4	Ação5	Ação6
e1	∅					
e2	FT	∅				
e3	FT	∅	∅			
e4	FT, NM	FT, NM	FT, NM	∅		
e5	FT, NM, PM	NM, PM	NM, PM	FT, PM	∅	
e6	NM, PM	FT, NM, PM	FT, NM, PM	FT, PM	FT	∅

Fonte: Adaptado de Sassi (2006).

- Função de Discernibilidade

A função de discernibilidade $F_s(B)$ é uma função booleana com m variáveis, que determina o conjunto mínimo de atributos necessários para diferenciar qualquer classe de equivalência das demais, definida na Equação (28) como:

$$F_s(\mathbf{a}_1^*, \mathbf{a}_2^*, \dots, \mathbf{a}_m^*) = \bigwedge \{ \bigvee m_D^*(i, j) \mid i, j = 1, 2, \dots, n, \quad m_D(i, j) \neq 0 \}.$$

$$\text{Sendo: } m_D^*(i, j) = \{ \mathbf{a}^* \mid \mathbf{a} \in m_D(i, j) \} \quad (28)$$

Utilizando o método de simplificação de expressões booleanas na função $F_s(B)$, obtém-se o conjunto de todos os implicantes primos dessa função, o qual determina os redutos de S . A simplificação é um processo de manipulação algébrica das funções lógicas com a finalidade de reduzir o número de variáveis e de operações necessárias para a sua realização.

A função de discernibilidade $F_s(B)$ é obtida da seguinte forma: para os atributos contidos dentro de cada célula da Matriz de Discernibilidade $MD(B)$ (Tabela 6), aplica-se o operador “soma”, “or” ou “ \vee ” e, entre as células dessa matriz, utiliza-se o operador “produto”, “and” ou “ \wedge ”, resultando em uma expressão booleana de “Produto da Soma”. A $F_s(B)$ é representada por:

$$F_s(B) = (FT) \wedge (FT) \wedge (FT \vee NM) \wedge (FT \vee NM) \wedge (FT \vee NM) \wedge (FT \vee NM \vee PM) \wedge (NM \vee PM) \wedge (NM \vee PM) \wedge (FT \vee PM) \wedge (NM \vee PM) \wedge (FT \vee NM \vee PM) \wedge (FT \vee NM \vee PM) \wedge (FT \vee PM) \wedge (FT)$$

Simplificando esta expressão, utilizando teoremas, propriedades e postulados da Álgebra Booleana, obtém-se a seguinte expressão minimizada:

$$F_s(B) = (FT \wedge (NM \vee PM) \wedge (FT \vee NM \vee PM))$$

Que ainda pode ser escrita na forma de “Soma do Produto”:

$$F_s(B) = (FT \wedge (NM \vee PM)).$$

Os redutos são:

$$RED(B) = \{ \text{Fechamento, Nova Mínima} \} \text{ e } \{ \text{Fechamento, Preço Máximo} \}.$$

A função de discernibilidade determinou o termo mínimo da função, ou seja, determinou o conjunto mínimo de atributos necessários para discernir as classes formadas por todas as classes de equivalência da relação $INDs(B)$.

2.4 Revisão Bibliométrica

Foi realizada uma revisão bibliométrica nas bases de dados: CAPES, Emerald, Google acadêmico, IEEEExplore, Proquest, Scielo, Science Direct, Science open e Scopus. O critério para a busca foram as palavras-chave: *absenteísmo*, *presenteísmo*, *mineração de dados* e *inteligência computacional*. A busca também foi realizada na língua inglesa com as palavras-chave: *absenteeism*, *presenteeism*, *data mining* e *computational intelligence*. Foram encontrados artigos apenas com as palavras combinadas: *absenteísmo* e *presenteísmo* e *absenteeism* e *presenteeism*. Apesar do *presenteísmo* estar relacionado com o *absenteísmo*, não é foco dessa pesquisa. Inicialmente foram coletados 200 artigos, que após a aplicação dos critérios de inclusão e exclusão, descritos a seguir, foram selecionados 100 artigos.

Os critérios de inclusão e exclusão para compor a revisão bibliométrica foram:

a) Critério de inclusão

- Artigos escritos em inglês, português e espanhol;
- Artigos relacionados com os cinco aspectos comuns:
 - i) Gerenciamento do *absenteísmo*;
 - ii) *Absenteísmo* relacionado à produtividade;
 - iii) *Absenteísmo* relacionado à saúde no trabalho;
 - iv) *Absenteísmo* relacionado à saúde pública;
 - v) *Absenteísmo* relacionado à segurança do trabalho.
- Artigos publicados a partir do ano de 1945 até janeiro do ano de 2019;
- Artigos com filiação em qualquer instituição de ensino superior, país e continente;

b) Critério de exclusão

- Artigos sobre *absenteísmo* escolar, acadêmico;
- Artigos não relacionados com os cinco aspectos comuns.

A seguir, são mostrados os resultados da bibliometria. Foram selecionados 100 artigos com origem em 33 países dos cinco continentes.

A Figura 19 mostra a distribuição dos artigos por país de origem.

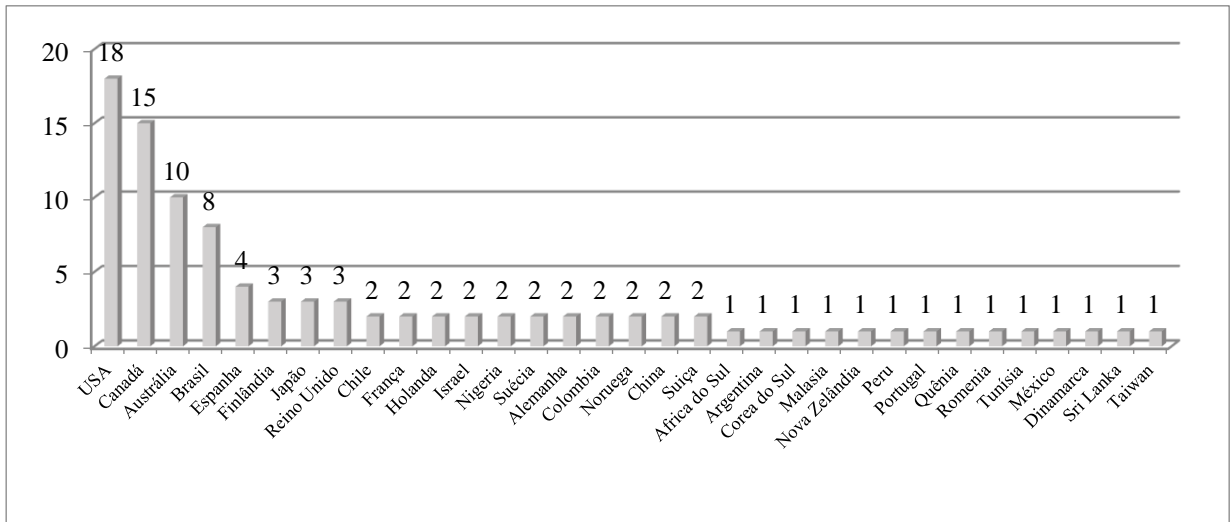


Figura 19 – Distribuição dos artigos por país de origem.

Observa-se que os países com o maior número de artigos que tratam de absenteísmo são: Estados Unidos da América, Canadá, Austrália e Brasil.

Percebe-se que os países mais desenvolvidos apresentam mais estudos sobre o absenteísmo, como USA e Canadá. A Tabela 7 mostra a distribuição dos artigos por continente.

Tabela 7 – Distribuição dos artigos por continente

Continente	Artigos	(%)
América	49	49%
Europa	24	24%
Oceania	11	11%
Ásia	11	11%
África	5	5%
Total	100	100%

O continente Americano e a Europa concentram a maior parte das pesquisas com 73% dos artigos relacionados com o absenteísmo.

A Figura 20 mostra a distribuição dos artigos pesquisados por aspectos comuns.

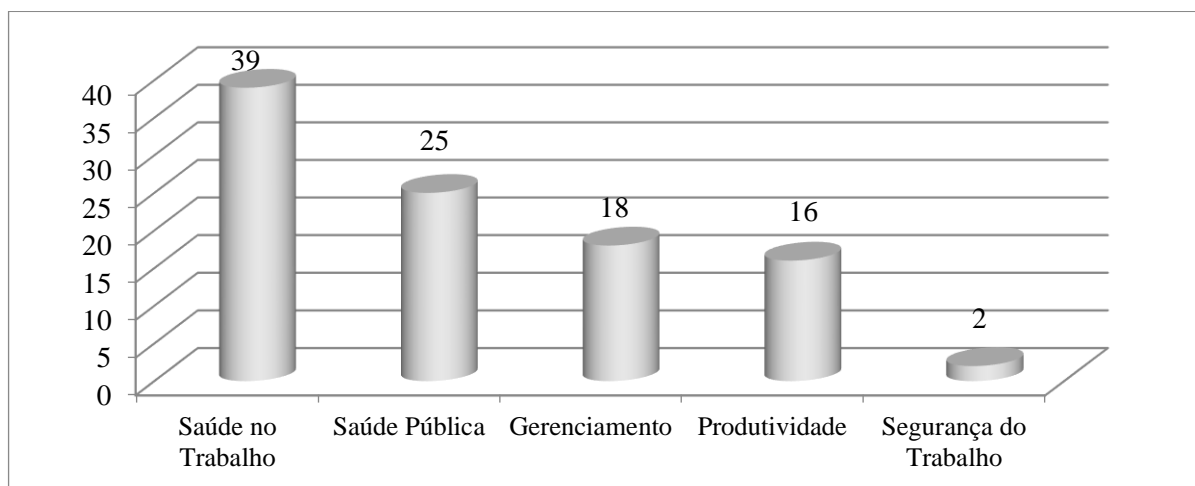


Figura 20 – Distribuição dos artigos pesquisados por aspecto comum.

Observa-se que 39% dos artigos (39) foram classificados com aspectos comuns à saúde no trabalho, saúde pública 25% dos artigos (25), gerenciamento 18% dos artigos (18), produtividade 16% dos artigos (16) e segurança do trabalho 2%. Percebe-se que a pesquisa se concentra na saúde no trabalho e na saúde pública juntas correspondem a 64% das pesquisas.

A Tabela 8 mostra a distribuição dos artigos por idioma.

Tabela 8 – Distribuição dos artigos por idioma

Idioma	Artigos	(%)
Inglês	86	86%
Português	7	7%
Espanhol	7	7%
Total	100	100%

Observa-se que os artigos escritos na língua inglesa predominam com 86% na língua portuguesa 7 artigos ou 7% e na língua espanhola 7 artigos ou 7% dos artigos pesquisados na bibliometria realizada.

Percebe-se que 73% dos artigos pesquisados concentram-se nos continentes Americano e Europeu, tendo maior classificação o aspecto comum em saúde no trabalho e tendo como idioma predominante a língua inglesa.

A Tabela 9 mostra a distribuição dos artigos por base de dados.

Tabela 9 – Distribuição dos artigos por base de dados

Base da dados	Artigos	(%)
Science Direct	37	37%
Scielo	15	15%
Google Acadêmico	14	14%
IEEE	12	12%
Proquest	7	7%
Emerald	5	5%
Scopus	4	4%
Capes	3	3%
Science Open	3	3%
Total	100	100%

A pesquisa abrange artigos publicados de 1945 a janeiro de 2019. A Tabela 10 mostra a distribuição dos artigos por período.

Tabela 10 – Distribuição dos artigos por período

Período	Artigos	(%)
Até 2000	5	5%
2001 - 2003	2	2%
2004 - 2006	3	3%
2007 - 2009	13	13%
2010 - 2012	27	27%
2013 - 2015	26	26%
2016 - 2018	22	22%
Até janeiro 2019	2	2%
Total	100	100%

Observa-se que existe um aumento nas pesquisas sobre absentéismo a partir de 2007 porque é considerado problema crítico para as empresas, uma vez que compromete o desempenho empresarial.

Foram apresentados os 100 artigos selecionados na revisão bibliométrica com o ano da publicação, título do artigo, periódico da publicação e referência. Esses artigos foram utilizados em sua maioria para compor o referencial teórico da pesquisa, que pode ser encontrado no Anexo A. Os artigos foram organizados em ordem cronológica decrescente.

Percebe-se na revisão bibliométrica que ao longo dos anos a pesquisa sobre o absentéismo migrou de aspectos comuns: produtividade e gerenciamento para aspectos comuns: saúde no trabalho, saúde pública e segurança do trabalho.

Do ponto de vista dos procedimentos metodológicos os artigos foram classificados em (GIL, 2010):

- Pesquisa bibliográfica ou revisão da literatura: elaboradas a partir de livros, artigos, periódicos, dissertações, teses e *websites*;
- Estudo de caso: envolve o estudo profundo e exaustivo de um ou poucos objetos de maneira que se permita o amplo e detalhado conhecimento;
- Pesquisa experimental: pesquisa em que se determina um objeto de estudo em que variáveis são selecionadas e definidas a forma de controle e de observação dos efeitos que as variáveis produzem no objeto de estudo;
- *Survey*: pesquisa que envolve o questionamento direto das pessoas cujo comportamento se deseja conhecer.

A Figura 21 mostra a classificação dos 100 artigos do ponto de vista dos procedimentos metodológicos.

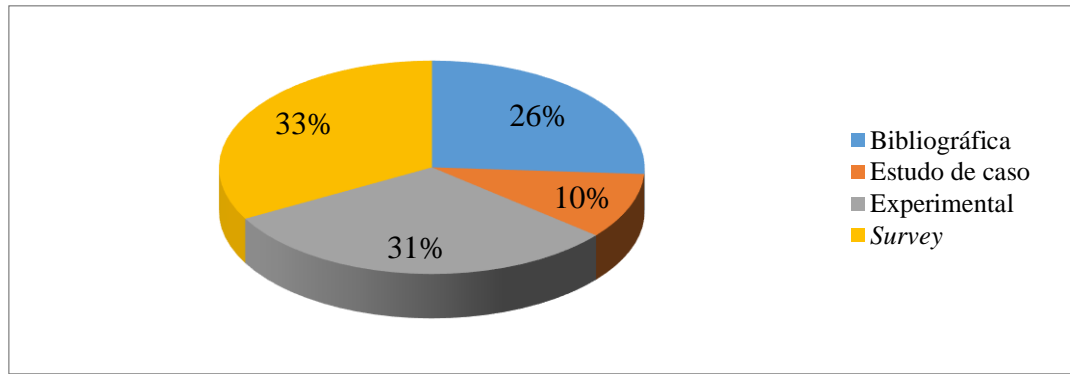


Figura 21 – Classificação dos artigos do ponto de vista dos procedimentos metodológicos.

Do ponto de vista da forma de abordagem do problema os artigos foram classificados conforme segue (GIL, 2010):

- Pesquisa qualitativa: considera que existe uma relação entre o mundo e o sujeito que não pode ser traduzida em números; a pesquisa é descritiva e o pesquisador tende a analisar seus dados indutivamente.

- Pesquisa quantitativa: considera que tudo é mensurável, o que significa traduzir opiniões e números em informações as quais serão classificadas e analisadas.

A Figura 22 mostra a classificação dos artigos do ponto de vista da forma de abordagem do problema.

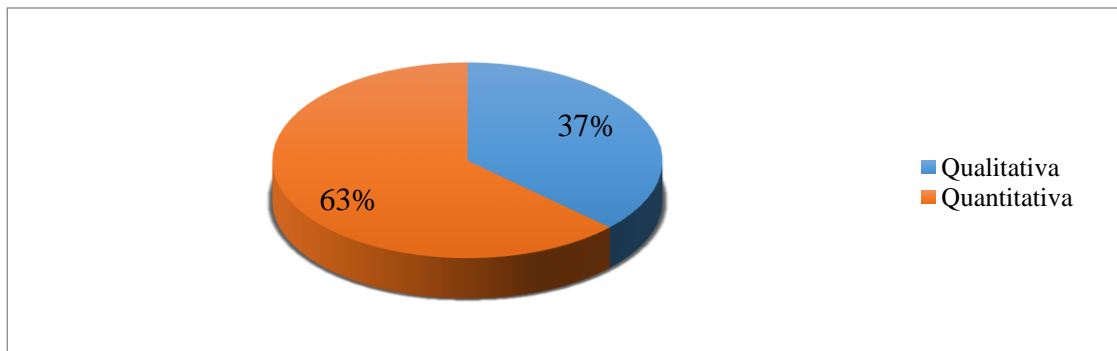


Figura 22 – Classificação dos artigos do ponto de vista da forma de abordagem do problema.

Tendo em vista o resultado da revisão bibliométrica, foram pesquisadas 100 obras que tratam o assunto sobre absenteísmo e presenteísmo, sendo que não foi encontrado nenhum artigo que relacione absenteísmo e presenteísmo com a utilização de mineração de dados ou inteligência computacional.

Embora o absenteísmo tenha sido pesquisado em nações, sociedades e classes sociais, os níveis de análise menores como nas empresas dos diversos segmentos ainda são pouco

investigados. Outro ponto observado é que apesar de que o impacto da influência social sobre o absenteísmo esteja estabelecido, pouco é sabido sobre a acumulação dinâmica e temporal dessa influência (RUSSO et al., 2013).

Após a revisão bibliométrica, confirmou-se que existem poucas informações e bases de dados sobre o assunto, dessa forma, a pesquisa proposta no trabalho possui relevância pois apresenta 120 meses de dados históricos sobre o absenteísmo, podendo contribuir com o entendimento da acumulação dinâmica e temporal sobre o absenteísmo, sendo dessa forma, uma lacuna da pesquisa.

Com os dados históricos surge a motivação para avançar com a previsão do absenteísmo e a identificação de tendências absenteístas com a finalidade de contribuir com o entendimento da acumulação dinâmica e temporal do absenteísmo.

Dessa forma, a inteligência computacional surge como alternativa de apoio ou substituição aos métodos tradicionais de processamento da informação. Assim sendo, espera-se prever o absenteísmo e identificar tendências absenteístas.

Como justificativa da pesquisa, é necessário citar o desafio de prever o absenteísmo e identificar tendências absenteístas, o que impulsiona a utilização da inteligência computacional, buscando a obtenção de resultados que possibilitem uma análise mais aprofundada e a consequente previsão do absenteísmo e a identificação de tendências absenteístas.

A previsão do absenteísmo pode auxiliar a tomada de decisão que antecede as operações logísticas, de forma a viabilizar alternativas de reposição de mão de obra ou priorização de operações de maior relevância, dessa maneira, minimizando impactos negativos nas operações logísticas. Já a identificação de tendência absenteísta pode auxiliar no tratamento e recuperação antes que o indivíduo se torne absenteísta.

Apresenta-se a seguir um levantamento das obras que abordaram a aplicação da computação e de técnicas da mineração de dados na área de gestão de pessoas e publicações recentes com estudos sobre o absenteísmo.

Rosas-Daniel et al. (2014) realizaram uma revisão de literatura do uso de *softwares* computacionais aplicado no gerenciamento de recursos humanos em problemas como previsão, avaliação e seleção de pessoal. Pombo, Araújo e Viana (2014), por outro lado, realizaram uma revisão sistemática dos sistemas de apoio à decisão clínica para o gerenciamento da dor. No caso de Jantan, Hamdan e Othman (2008), foram apresentadas potenciais técnicas inteligentes aplicadas no suporte à tomada de decisão na área de gestão de pessoas.

Stavrou, Charalambous e Spiliotis (2007) e Ning (2010) aplicaram uma rede neural

artificial na alocação e gestão de recursos humanos; e Li, Xu e Meng (2009) desenvolveram um modelo de predição de risco dos recursos humanos alocados em uma empresa da construção civil baseado na técnica *Support Vector Machine* (SVM). Mei (2009) apresentou um modelo de abastecimento de recursos humanos baseado em uma rede neural artificial do tipo *Radial Basis Function* (RBF) com a Matriz de Transferência de Markov.

Rodrigues, Oliveira e Sa (2011) aplicaram um Modelo Multicritério para alocar recursos humanos em programas de assistência na comunidade; e Wang e Jiang (2010) apresentaram um estudo da avaliação humana em empresas baseado em redes neurais artificiais. Yannibelli e Amandi (2011) maximizaram a eficácia dos recursos humanos no contexto do projeto de agendamento por meio de algoritmo evolutivo; e Jantan, Hamdan e Othman (2010) também aplicaram árvores de decisão na predição de talentos.

Strohmeier e Piazza (2013), Jantan, Hamdan e Othman (2009) e Sadath (2013) aplicaram, respectivamente, a mineração de dados no gerenciamento de recursos humanos, no gerenciamento de talentos e como ferramenta para gerenciamento do conhecimento em recursos humanos.

Parker, Morgeson e Johns (2017) aplicaram técnicas de agrupamentos para verificar as perspectivas quanto a custos e benefícios de diferentes tipos de projeto de trabalho, dessa forma encontraram um conjunto diversificado de resultados.

O trabalho desenvolvido por Lopes et al. (2017) apresenta um estudo sobre o fenômeno presenteísmo fundamentando-se na construção de um mapa sistêmico, que inclui as principais variáveis que circundam o tema, o estudo apresenta um modelo matemático *fuzzy* para mensurar as relações entre elas e automatizar tal representação sistêmica. Para tanto, utilizou-se técnica de análise e julgamento para a tomada de decisão, gerando uma leitura quantitativa para as variáveis qualitativas representadas no mapa sistêmico.

Os Materiais e Métodos são apresentados no próximo capítulo.

3. MATERIAIS E MÉTODOS

Neste capítulo são apresentados os materiais e métodos utilizados no trabalho e a descrição detalhada dos experimentos.

3.1 Caracterização metodológica

A metodologia adotada para o desenvolvimento deste trabalho foi definida como experimental (GIL, 2010). A pesquisa bibliográfica foi elaborada a partir de artigos de periódicos, livros, teses, dissertações e materiais disponibilizados em bases de dados de *websites* como Scielo, IEEEExplore, Scopus, Emerald, Science Direct e de congressos com conteúdo sobre absentismo, KDD, mineração de dados, técnicas da inteligência artificial, redes neurais artificiais, *Multilayer Perceptron*, rede SOM e a teoria dos *Rough Sets*.

Esse trabalho completo contém 234 referências bibliográficas incluídas as referências da revisão bibliométrica (Anexo A) e são apresentadas na Tabela 11, considerando o período, quantidade de publicações e a representatividade em porcentagem.

Tabela 11 – Período e quantidade de publicações que compõem o estudo

Período	Publicações	Porcentagem
Até o ano de 1999	17	7,26%
2000 a 2006	21	8,97%
2007 a 2013	107	45,73%
2014 a 2019	89	38,03%
Total	234	100,00%

A Tabela 12 apresenta a distribuição por tipo de publicação. As citações de livros e conferências/simpósios foram incluídas no trabalho devido à relevância para o desenvolvimento do trabalho.

Tabela 12 – Distribuição por tipo de publicação

Tipo de publicação	Quantidade	Porcentagem
Periódicos	182	77,78%
Conferências/Simpósios	15	6,41%
Livros	26	11,11%
Dissertações	4	1,71%
Teses	2	0,85%
<i>Websites</i>	5	2,14%
Total	234	100,00%

A pesquisa experimental representa o melhor exemplo de pesquisa científica. Essencialmente, consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto, é um tipo de pesquisa que visa elucidar relações de causa e efeito entre variáveis (GIL, 2010; APPOLINÁRIO, 2011).

Quanto à natureza, a pesquisa foi classificada como aplicada, visto que tem como objetivo gerar conhecimento para a resolução de problemas ou necessidades concretas e imediatas, permitindo uma aplicação prática, muitas vezes, nessa modalidade de pesquisa, os problemas emergem do contexto profissional e podem ser sugeridos pela instituição para que o pesquisador solucione uma situação-problema. Quanto à abordagem foi classificada como quantitativa (RODRIGUES e LIMENA, 2006; GIL, 2010; YIN, 2015; APPOLINÁRIO, 2011).

De acordo com Rodrigues e Limena (2006) a pesquisa quantitativa é compreendida quando a abordagem está relacionada à quantificação, análise e interpretação de dados obtidos mediante pesquisa voltada para análise e a interpretação dos resultados.

3.2 Caracterização da empresa

Uma empresa de logística localizada na cidade de São Paulo foi o objeto de estudo deste trabalho. Trata-se de uma empresa pública com atuação em todo o território nacional e conta

com cerca de cem mil empregados. A motivação para essa escolha foi a disponibilização no acesso às informações e apoio técnico para coleta dos dados. A empresa por meio de sua Universidade Corporativa também apoiou a pesquisa na forma de uma bolsa de estudo cedida ao autor deste trabalho.

3.3 Base de dados, *softwares* e plataforma de experimentos

A base de dados é oriunda do seguinte artigo “Martiniano, A.; Ferreira, R. P.; Sassi, R. J.; Affonso, C. *Application of a neuro fuzzy network in prediction of absenteeism at work*. In *Information Systems and Technologies (CISTI)*, 7th Iberian Conference on (p. 1-4). IEEE, June, 2012”.

Após a publicação do artigo, a base foi cedida para o repositório da UCI *Machine Learning Repository* (<http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>), já que bases de dados deste tipo são raras para realização de pesquisas e, desta forma, contribuir com os estudos relacionados ao tema deste trabalho. Até o momento de fechamento do texto (12/03/2019) foram realizados 134.478 acessos nesse repositório no *website* com 6 citações no google acadêmico e cerca de 40 contatos via *e-mail* de 10 países: África do Sul, Argélia, Canadá, Estados Unidos, Portugal, Finlândia, Coreia do Sul, Irã, Índia e Japão dos continentes africano, americano, europeu e asiático.

Os interessados na base de dados sobre o absentéismo foram alunos de graduação, alunos de mestrado, alunos de doutorado, doutores, professores, pesquisadores e profissionais de tecnologia da informação.

A base de dados utilizada é considerada de fonte secundária não sendo necessária a submissão a um comitê de ética em pesquisa (SILVA e CAVEIÃO, 2016). Segundo Mattar (2014), dados secundários são aqueles que já foram coletados, tabulados, ordenados e até analisados, ou seja, já existem e já foram publicados com propósitos de atender às necessidades de uma pesquisa.

Trata-se de uma base de dados com informações de pessoas que pela própria natureza tem diferentes formas de agir e de pensar diante de uma mesma situação o que pode trazer ambiguidade e incoerência nos dados coletados, dessa forma, criando ruído nos dados.

A composição da base de dados está descrita no Quadro 2 a seguir. A base de dados completa possui 50 atributos com 2403 registros.

Quadro 2 – Descrição dos 50 atributos que compõem a base de dados

Atributos	Descrição
1	Corresponde ao código individual
2	Corresponde aos motivos do absentéismo (Detalhado nas Tabelas 13 e 14)
3	Corresponde ao tempo de permanência do indivíduo na base de dados em anos
4	Corresponde ao mês do registro
5	Corresponde ao ano do registro
6	Corresponde ao dia da semana
7	Corresponde à estação do ano do registro
8	Corresponde ao tempo de afastamento em dias
9	Corresponde ao tempo de afastamento em horas
10	Corresponde ao tempo de afastamento em dias e horas medidos e transformados em horas
11	Corresponde a quantidade de licenças médicas registradas por mês
12	Corresponde ao valor em reais do transporte mensal utilizado da residência ao trabalho e do trabalho para a residência
13	Corresponde à distância da residência para o trabalho em quilômetros
14	Corresponde ao valor em reais do transporte mensal multiplicado pela distância da residência para o trabalho em quilômetros dividida por mil
15	Corresponde ao tempo de serviço em anos completos
16	Corresponde ao tempo de serviço em meses completos
17	Corresponde ao tempo de serviço em dias
18	Corresponde à idade em anos completos
19	Corresponde à idade em meses completos
20	Corresponde à idade em dias
21	Corresponde à carga de trabalho anual
22	Corresponde à carga de trabalho mensal
23	Corresponde à carga média de trabalho diária
24	Corresponde a meta atingida
25	Corresponde a escolaridade
26	Corresponde ao número de filhos

27	Corresponde ao etilismo social
28	Corresponde ao tabagismo social
29	Corresponde a animais de estimação
30	Corresponde ao peso
31	Corresponde à altura
32	Corresponde ao índice de massa corporal (variável contínua)
33	Corresponde ao índice de massa corporal (variável categórica)
34	Corresponde ao calçado
35	Corresponde tronco
36	Corresponde a cintura
37	Corresponde as despesas médico-hospitalares e odontológicas por idade em percentagem
38	Corresponde a remuneração mensal média
39	Corresponde a remuneração média por hora
40	Corresponde a remuneração média por hora de absentéismo
41	Corresponde a classificação do indivíduo absenteísta (variável categórica)
42	Corresponde a classificação do indivíduo absenteísta (variável contínua)
43	Corresponde a frequência das licenças médicas por indivíduo
44	Corresponde ao absentéismo mensal medido em horas
45	Corresponde ao absentéismo bimestral medido em horas
46	Corresponde ao absentéismo trimestral medido em horas
47	Corresponde ao absentéismo quadrimestral medido em horas
48	Corresponde ao absentéismo semestral medido em horas
49	Corresponde ao absentéismo anual medido em horas
50	Corresponde ao absentéismo bianual medido em horas.

O segundo atributo apresenta os motivos do absenteísmo estratificados em vinte e sete categorias, sendo vinte e uma categorias segundo a Classificação Internacional de Doenças (CID) e seis categorias sem CID, conforme as Tabelas 13 e 14, juntas formam um atributo.

Tabela 13 – Classificação Internacional de Doenças (CID)

Capítulo	Descrição
1	Algumas doenças infecciosas e parasitárias (A00-B99)
2	Neoplasias [tumores] (C00-D48)
3	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários (D50-D89)
4	Doenças endócrinas, nutricionais e metabólicas (E00-E90)
5	Transtornos mentais e comportamentais (F00-F99)
6	Doenças do sistema nervoso (G00-G99)
7	Doenças do olho e anexos (H00-H59)
8	Doenças do ouvido e da apófise mastóide (H60-H95)
9	Doenças do aparelho circulatório (I00-I99)
10	Doenças do aparelho respiratório (J00-J99)
11	Doenças do aparelho digestivo (K00-K93)
12	Doenças da pele e do tecido subcutâneo (L00-L99)
13	Doenças do sistema osteomuscular e do tecido conjuntivo (M00-M99)
14	Doenças do aparelho geniturinário (N00-N99)
15	Gravidez, parto e puerpério (O00-O99)
16	Algumas afecções originadas no período perinatal (P00-P96)
17	Malformações congênicas, deformidades e anomalias cromossômicas (Q00-Q99)
18	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte (R00-R99)
19	Lesões, envenenamento e algumas outras conseqüências de causas externas (S00-T98)
20	Causas externas de morbidade e de mortalidade (V01-Y98)
21	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde (Z00-Z99)

Fonte: Adaptado de WHO (2019).

A Classificação Internacional de Doenças e Problemas Relacionados à Saúde também conhecida como Classificação Internacional de Doenças (CID 10) é publicada pela Organização Mundial de Saúde (OMS) do inglês *World Health Organization* (WHO) e visa padronizar a codificação de doenças e outros problemas relacionados à saúde.

A CID 10 fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças. A cada estado de saúde é atribuída uma categoria única à qual corresponde um código CID 10 (WHO, 2019).

A Tabela 14 apresenta os seis motivos (categorias) de absenteísmo não classificadas pelo CID.

Tabela 14 – Lista dos seis motivos (categorias) sem CID

Motivo/Categoria	Descrição
22	Acompanhante
23	Consulta médica
24	Doador de sangue
25	Exame
26	Fisioterapia
27	Consulta odontológica

Os motivos do absenteísmo sem CID foram incluídos na base de dados uma vez que outros motivos do absenteísmo além dos classificados pelo CID também são relevantes para a composição do trabalho.

As categorias de absenteísmo não classificadas pelo CID são descritas a seguir:

- Acompanhante: Absenteísmo relacionado ao acompanhamento familiar a serviços de saúde de genitores idosos ou progenitores, cônjuge ou filhos menores de idade devidamente comprovada. Normalmente previsto em acordo coletivo de trabalho.

- Consulta médica: Absenteísmo de curta duração (menor que um dia ou menor que oito horas) para realização de consultas médicas que não apresentam a notação de CID;

- Doador de sangue: O empregado poderá deixar de comparecer ao serviço sem prejuízo de salário, por um dia, em cada 12 (doze) meses de trabalho, em caso de doação voluntária de sangue devidamente comprovada. Trata-se de absenteísmo amparado pela LEI 1075/1950, Inciso IV (SANTOS, 2015);

- Exame: Absenteísmo de curta duração para realização de exames diversos sem a notação de CID;

- Fisioterapia: Absenteísmo de curta duração para realização de fisioterapia;

- Consulta odontológica: Absenteísmo de curta duração para realização de tratamento odontológico.

A Tabela 15, apresenta os atributos que compõem a base de dados os valores mínimos e máximos por atributo que compõem a base de dados utilizadas nos experimentos computacionais.

Tabela 15 – Lista de atributos que compõem a base de dados

Item	Atributos	Mínimo	Máximo
1	Código Individual	1	39
2	CID / Motivo da ausência	1	27
3	Tempo de permanência na base	1,6	10
4	Mês	1	12
5	Ano	2008	2017
6	Dia da Semana	2	6
7	Estação do ano	1	4
8	Tempo de afastamento em dias	0	15
9	Tempo de afastamento em horas	0	7
10	Tempo de afastamento (dia+horas) em horas	1	120
11	Quantidade de atestados por mês	4	48
12	Valor do transporte para o trabalho mensal em reais (R\$)	118,8	388,7
13	Distância da residência para o trabalho (Km)	4,9	133
14	(Valor_Transp x Distância)/1000	1,22	51,07
15	Tempo de serviço (Em anos)	0	57
16	Tempo de serviço (Em meses)	6	695
17	Tempo de serviço (Em dias)	196	21160
18	Idade (Em anos)	24	73
19	Idade (Em meses)	296	883
20	Idade (Em dias)	9039	26882
21	Carga de trabalho/ano	62450220	80003120
22	Carga de trabalho/mês	4236560	10350518
23	Carga de trabalho (Média/dia)	192571	470478
24	Meta atingida	81,4	100
25	Escolaridade	1	4
26	Número de filhos	0	4
27	Etilista	0	1
28	Tabagista	0	1
29	Animais Estimação	0	8
30	Peso	56	110
31	Altura	1,63	1,96
32	Índice de massa corporal_IMC_Real	19,2	38
33	Índice de massa corporal_IMC_Categoria	1	4
34	Pé	38	43
35	Tronco	3	6
36	Cintura	38	58
37	Despesas Médico-hospitalares e odontológica por idade (%)	2,5	47,4
38	Remuneração	2912	7971
39	Remuneração_p/h	18,2	49,82
40	Remuneração_p/h_ABS	18	4384
41	Class_Real_0_to_1	0	1
42	Class_INT_1P_to_6A	1	6
43	Frequência Atestado Médico	0	1196
44	Absenteísmo_Mensal	41	417
45	Absenteísmo_Bimestral	124	707
46	Absenteísmo_Trimestral	263	1124
47	Absenteísmo_Quadrimestral	376	1259
48	Absenteísmo_Semestral	618	1597
49	Absenteísmo_Anual	1245	2812
50	Absenteísmo_Bianual	2880	5202

O Quadro 3 mostra os *softwares* que foram utilizados no trabalho, utilização, descrição e *website*.

Quadro 3 – *Softwares* utilizados no trabalho

<i>Software</i>	<i>Utilização</i>	<i>Descrição</i>	<i>Website</i>
MBP	Utilizado nos experimentos computacionais para a aplicação da MLP	O <i>Multiple Back-Propagation</i> é um <i>software</i> livre para desenvolvimento de experimentos computacionais com redes neurais artificiais (MBP, 2019).	http://mbp.sourceforge.net/
SOMine	Utilizado nos experimentos computacionais para aplicação da rede SOM	É considerado um <i>software</i> amigável, flexível e poderoso (KOHONEN, 2001; VISCOVERY, 2019). Trata-se de um <i>software</i> comercial.	https://www.viscovery.net/
Rosetta	Utilizado nos experimentos computacionais para aplicação dos <i>Rough Sets</i> . No caso da redução de atributos disponibiliza dois tipos de algoritmos para realizar a redução, o Algoritmo Genético e o Algoritmo de Johnson.	Foi desenvolvido com esforço cooperativo envolvendo o Grupo de Sistemas de Conhecimento (NTNU), na Noruega, e o Grupo de Lógica da Universidade de Varsóvia, na Polônia. Trata-se de um <i>software</i> livre, pode ser baixado gratuitamente de acordo com a licença (NTNU, 2019).	https://www.ntnu.edu/idi/

A plataforma de *hardware* utilizada nos experimentos foi um microcomputador com processador Intel® Core™ CPU de 2.00 GHz, com 4 GB de memória RAM e 1 TB de disco rígido, sistema operacional de 64 *bits* e Windows 10, Fabricante Samsung.

Os parâmetros usados na RNA foram os seguintes: neurônios na camada de entrada, número de camadas ocultas, número de neurônios em cada camada oculta, número máximo de iterações (critério de parada), taxa de aprendizado e o fator de momento.

Os parâmetros usados na rede SOM foram os seguintes: número de neurônios, vizinhança topológica, função de vizinhança topológica, número de iterações e taxa de aprendizado. No caso dos *Rough Sets* a redução de atributos foi realizada pela aplicação das técnicas Algoritmo Genético e Algoritmo de Johnson presentes no *software* Rosetta.

3.4 Condução dos experimentos computacionais

Os experimentos computacionais foram divididos em duas Fases distintas denominadas Previsão do Absenteísmo contendo duas etapas e Identificação de Tendências Absenteístas contendo uma etapa.

Apresenta-se na Figura 23 as fases e etapas dos experimentos computacionais.

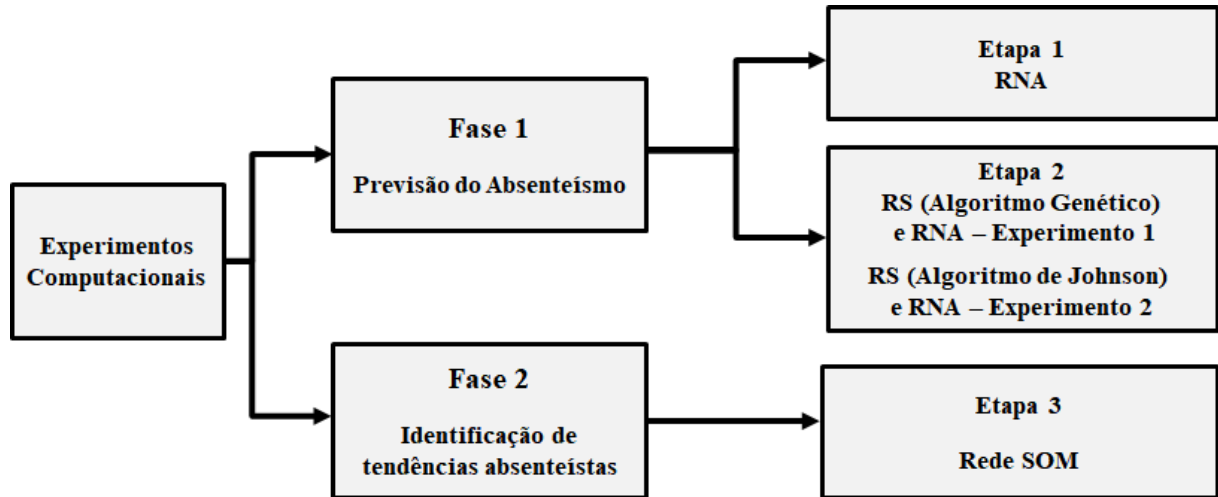


Figura 23 – Fases e etapas dos experimentos computacionais.

A seguir descreve-se as fases e as etapas dos experimentos computacionais.

- **Fase 1: Previsão do Absenteísmo** - Esta fase foi dividida em duas etapas (1 e 2).

Na Etapa 1 denominada RNA aplicou-se uma RNA do tipo MLP na previsão do absenteísmo, considerando os seguintes períodos: mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual. Nesta etapa foi realizado o treinamento da MLP.

Na mesma etapa para teste foram considerados os seguintes períodos: mensal, trimestral, quadrimestral, de forma que os resultados pudessem ser comparados por período. A validação dos resultados nesta Fase 1 foi feita por meio da verificação do erro experimental. Para a realização dos experimentos foram utilizadas três topologias diferentes de RNA.

Na Etapa 2 denominada RS + RNA aplicou-se *Rough Sets* na redução de atributos e em seguida uma MLP foi utilizada na previsão do absenteísmo. Foram realizados dois experimentos:

-**Experimento 1:** redução de atributos com algoritmo genético. Foram considerados os seguintes períodos para previsão do absenteísmo: mensal, trimestral e quadrimestral. Nesta etapa foi realizado o treinamento e Teste da MLP. O reduto selecionado foi o com o menor número de atributos ou maior redução. A validação dos resultados foi feita por meio da verificação do erro experimental da MLP.

-**Experimento 2:** redução de atributos com algoritmo de Johnson. Foram considerados os seguintes períodos para previsão do absenteísmo: mensal, trimestral e quadrimestral. Nesta etapa foi realizado o treinamento e Teste da MLP. O algoritmo de Johnson gerou um único reduto com 19 atributos. Os experimentos com o *Rough Sets* e RNA foram realizados com os

softwares Rosetta e MBP respectivamente. A validação dos resultados foi feita por meio da verificação do erro experimental da MLP.

Vale ressaltar que na Etapa 1 os experimentos computacionais foram realizados com a base de dados completa e na Etapa 2 com a base de dados reduzida por *Rough Sets*.

Apresenta-se as etapas 1 e 2 dos experimentos computacionais que compõem a Fase 1, conforme Figura 24.

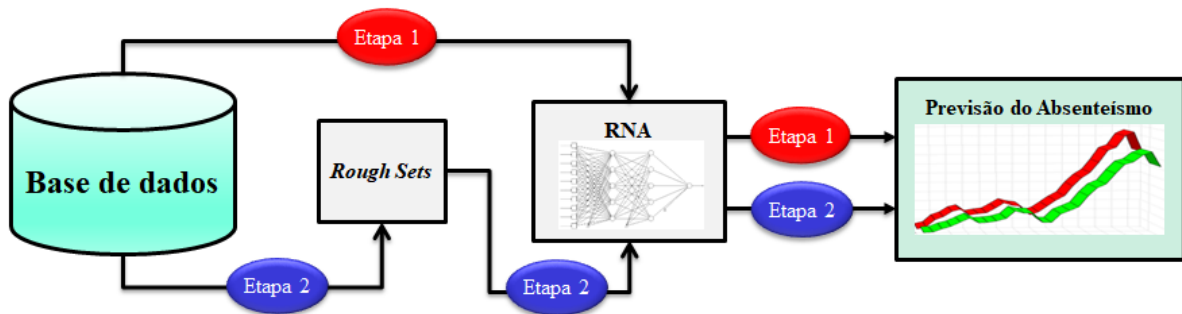


Figura 24 – Etapas 1 e 2 dos experimentos computacionais que compõem a Fase 1.

- **Fase 2: Identificação de tendências absenteístas** - Na Etapa 3 foi utilizada a rede SOM para identificação de tendências absenteístas nos agrupamentos gerados.

Descreve-se a seguir de forma detalhada a Etapa 3 da fase 2 dos experimentos computacionais conforme Figura 25.

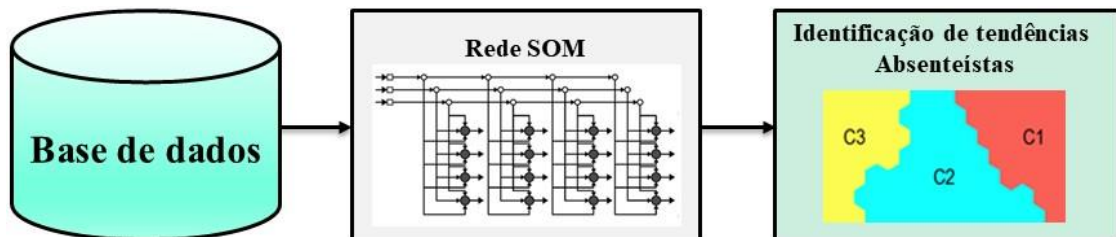


Figura 25 – Etapa 3 dos experimentos computacionais que compõem a Fase 2.

A avaliação dos resultados da rede SOM foi feita por meio da verificação do erro de quantização.

No próximo capítulo serão apresentados os resultados dos experimentos computacionais realizados.

4. APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo os resultados dos experimentos computacionais são apresentados e discutidos.

Como descrito no capítulo 3, a realização dos experimentos foi dividida em duas fases com três etapas distintas:

- **Fase 1:** Previsão do Absenteísmo dividida em duas etapas (1 e 2).
 - **Etapa 1 (RNA):** Aplicação da MLP na previsão do absenteísmo mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual na fase de Treinamento e mensal, trimestral e quadrimestral na fase de Teste.
 - **Etapa 2 (RS + RNA):** Aplicação dos *Rough Sets* e da MLP na previsão do absenteísmo dividido em dois experimentos.
 - a) Experimento 1 - *Rough Sets* usando algoritmo genético na redução de atributos para previsão do absenteísmo mensal, trimestral e quadrimestral nas fases de Treinamento e Teste.
 - b) Experimento 2 - *Rough Sets* usando algoritmo Johnson na redução de atributos para previsão do absenteísmo mensal, trimestral e quadrimestral nas fases de Treinamento e Teste.
- **Fase 2:** Identificação de tendências absenteístas
 - **Etapa 3 (SOM):** Aplicação da rede SOM na Identificação de tendências absenteístas.

4.1 Experimentos computacionais

O Quadro 4 apresenta as fases dos experimentos computacionais.

Quadro 4 – Fases dos experimentos computacionais

Fase 1: Previsão do Absenteísmo	Etapa 1 (RNA): Aplicação da MLP na previsão do absenteísmo.
	Etapa 2 (RS + RNA): Aplicação de <i>Rough Sets</i> e da MLP na previsão do absenteísmo.
Fase 2: Identificação de tendências absenteístas	Etapa 3 (SOM): Aplicação da rede SOM na Identificação de tendências absenteístas.

Os experimentos computacionais da Fase 1, considerando os resultados das fases de treinamento e teste são apresentados a seguir.

4.2 Fase 1 - Etapa 1 (RNA): Aplicação de uma RNA na previsão do absenteísmo (Treinamento)

Os experimentos com a MLP foram realizados com o objetivo de prever o absenteísmo mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual. A Figura 26 ilustra a realização dos experimentos da Fase 1, Etapa 1 da fase de Treinamento.

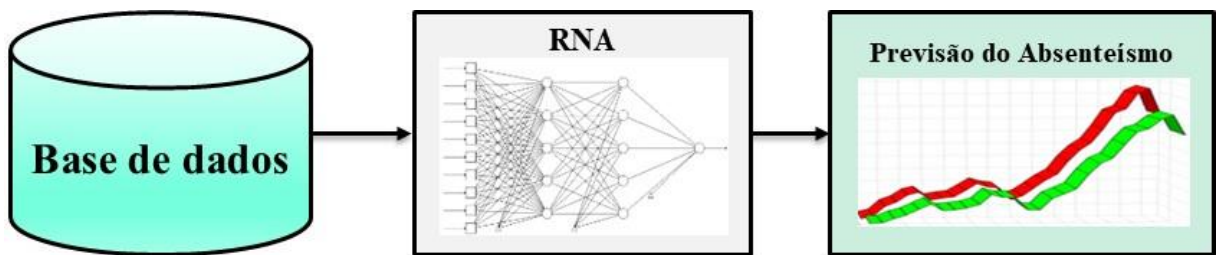


Figura 26 – Realização dos experimentos da Fase 1, Etapa 1 da fase de Treinamento.

Dos 50 atributos da base de dados, 43 foram utilizados como as entradas da MLP e 7 como as saídas: absenteísmo medido em horas mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual.

A Tabela 16 apresenta os 43 atributos de entrada da MLP.

Tabela 16 – Atributos de entrada da MLP (43 atributos)

Item	Atributos	Item	Atributos
1	Código Individual	23	Carga de trabalho (Média/dia)
2	CID / Motivo da ausência	24	Meta atingida
3	Tempo de permanência na base	25	Escolaridade
4	Mês	26	Numero de filhos
5	Ano	27	Etilista
6	Dia da Semana	28	Tabagista
7	Estação do ano	29	Animais Estimação
8	Tempo de afastamento em dias	30	Peso
9	Tempo de afastamento em horas	31	Altura
10	Tempo de afastamento (dia+horas) em horas	32	Índice de massa corporal_IMC_Real
11	Quantidade de atestados por mês	33	Índice de massa corporal_IMC_Categoria
12	Valor do transporte para o trabalho mensal em reais (R\$)	34	Pé
13	Distância da residência para o trabalho (Km)	35	Tronco
14	(Valor_Transp x Distância)/1000	36	Cintura
15	Tempo de serviço (Em anos)	37	Despesas Médico-hospitalares e odontológica por idade (%)
16	Tempo de serviço (Em meses)	38	Remuneração
17	Tempo de serviço (Em dias)	39	Remuneração_p/h
18	Idade (Em anos)	40	Remuneração_p/h_ABS
19	Idade (Em meses)	41	Class_Real_0_to_1
20	Idade (Em dias)	42	Class_INT_1P_to_6A
21	Carga de trabalho/ano	43	Frequência Atestado Médico
22	Carga de trabalho/mês		

A Tabela 17 apresenta os parâmetros utilizados na MLP com 43 neurônios de entrada.

Tabela 17 – Parâmetros utilizados na MLP (43 neurônios de entrada)

Parâmetros	Valor
Número de neurônios de entrada	43
Número de camadas ocultas	2
Número de neurônios nas camadas ocultas	10
Taxa inicial de aprendizado	0,7
Decaimento a cada 500 épocas	1%
Fator de momento inicial	0,7
Decaimento a cada 1000 épocas	1%

A função de ativação dos neurônios das camadas ocultas e das saídas da RNA foi a sigmoide, os pesos foram inicializados aleatoriamente entre -1 e 1, o *bias* fixado em -1, o critério de parada foi o erro quadrático médio na fase de treinamento menor que 0,01 ou 8.500 épocas.

A Figura 27 ilustra a topologia da MLP utilizada no experimento Etapa 1 fase de Treinamento.

O treinamento da MLP foi sequencial com 8.500 épocas e o tempo de processamento foi de 17 minutos e 13 segundos. O erro quadrático médio na fase de treinamento foi igual a 0,0346047245. As saídas foram o absenteísmo medido em horas mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual.

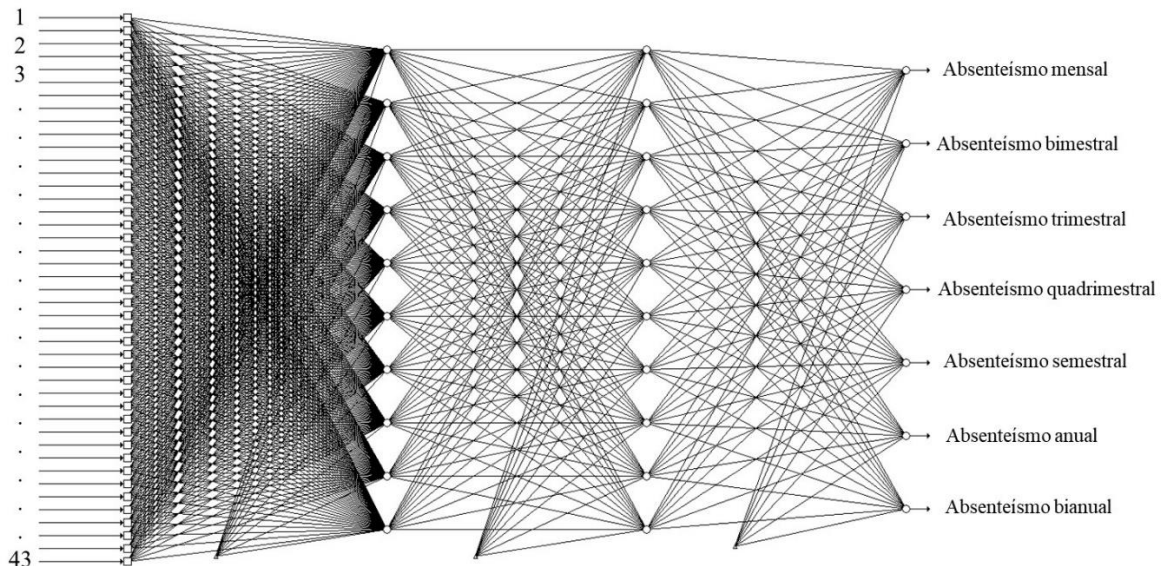


Figura 27 – Topologia da MLP utilizada no experimento.

A Figura 28 mostra a fase de treinamento da MLP para a previsão do absenteísmo **mensal** medido em horas.

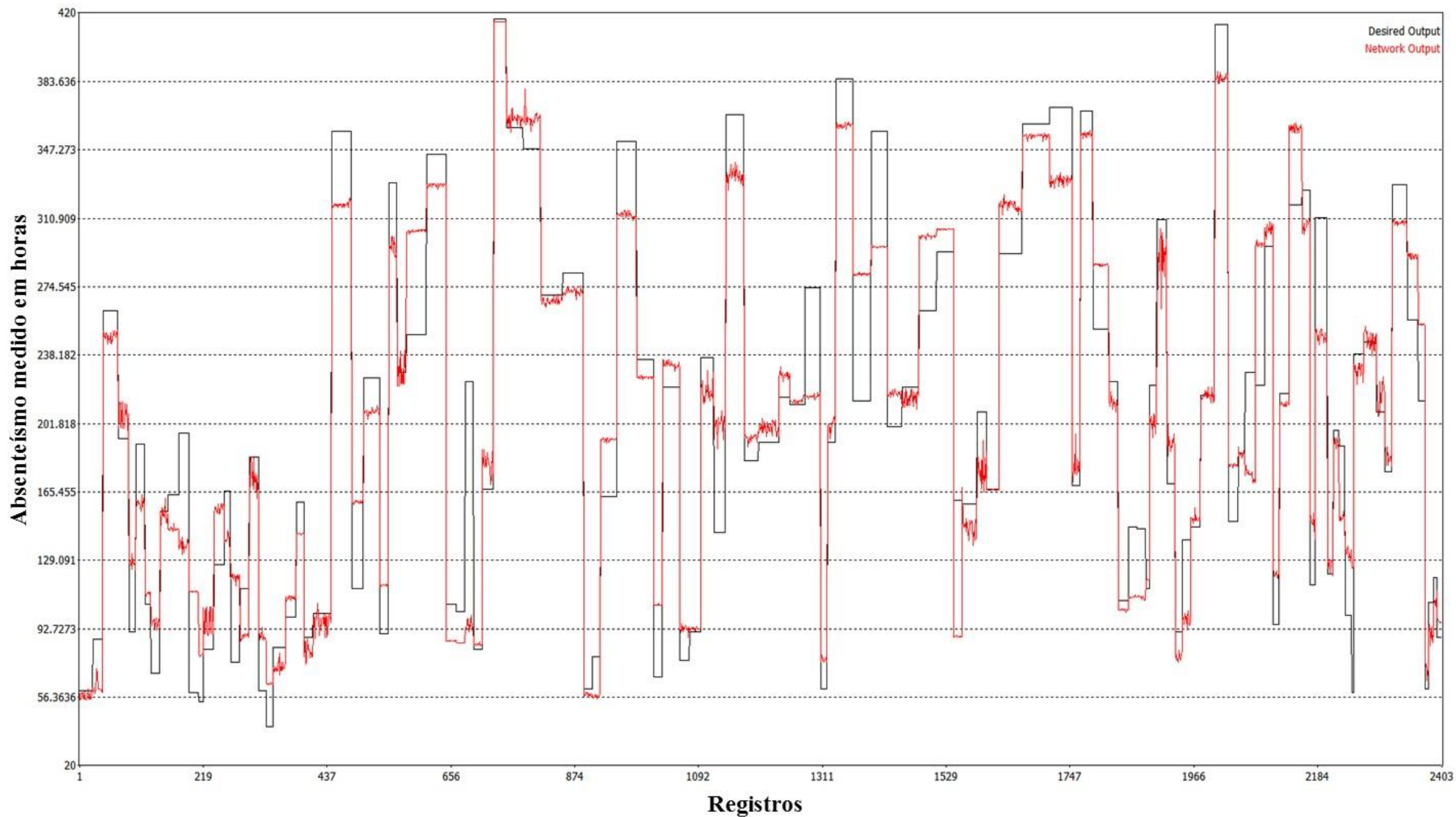


Figura 28 – Fase de treinamento da MLP para a previsão do absenteísmo **mensal** medido em horas.

O eixo x possui o intervalo de 1 a 2.403 e corresponde aos registros completos da base de dados. A curva na cor preta é a saída desejada e a curva na cor vermelha a saída da MLP.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados, tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245, percebe-se que nos intervalos de 656 a 874 e de 1529 a 1747 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada. Vale ressaltar que se trata de uma base de dados com informações de pessoas que pela própria natureza tem diferentes formas de agir e de pensar diante de uma mesma situação o que pode trazer ambiguidade e incoerência aos dados coletados (BECKER, 2008).

A Figura 29 mostra a fase de treinamento da MLP para a previsão do **absenteísmo bimestral** medido em horas.

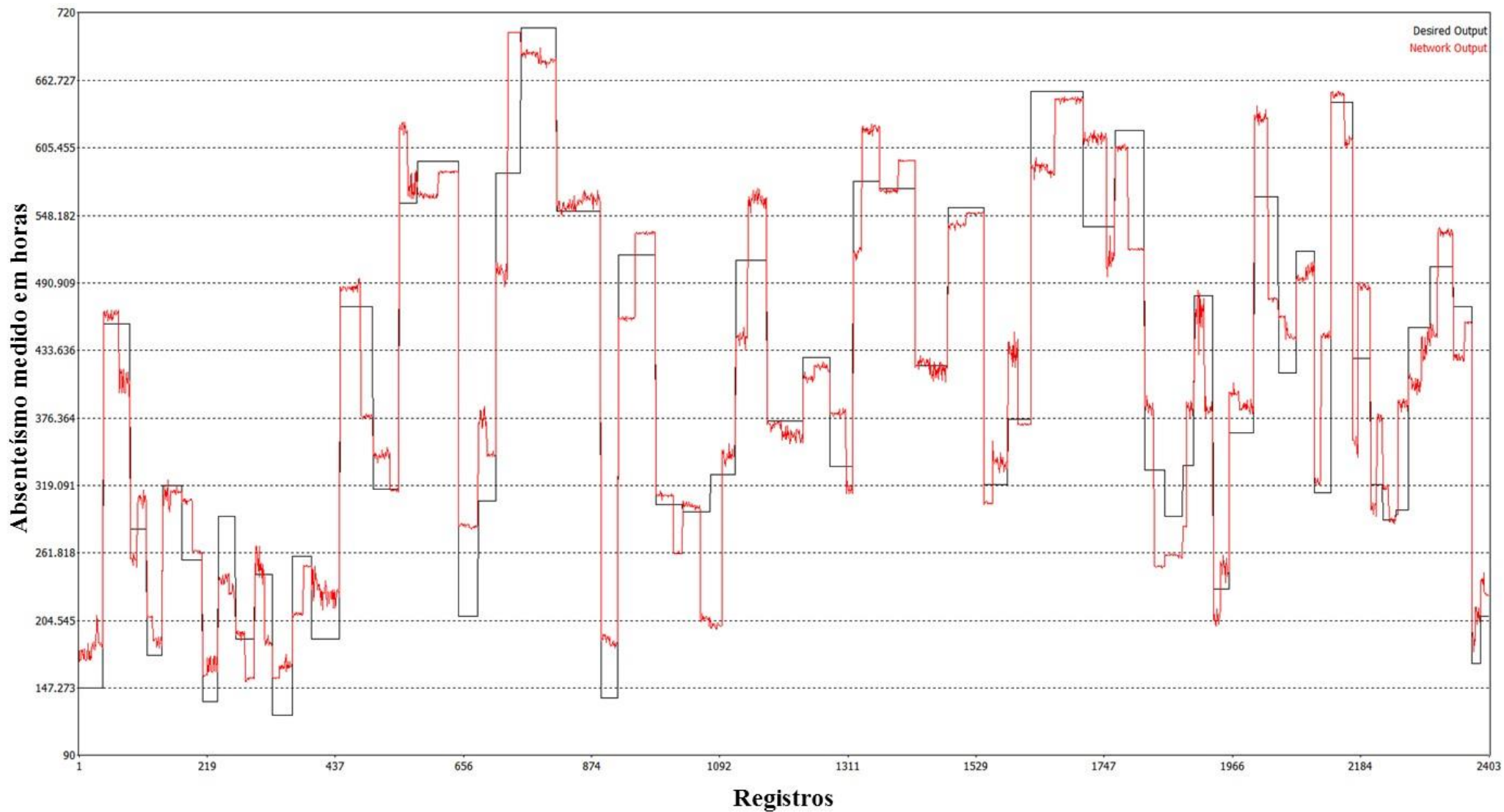


Figura 29 – Fase de treinamento da MLP para a previsão do absenteísmo **bimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245. Percebe-se que nos intervalos de 656 a 874 e de 1092 a 1311 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 30 mostra a fase de treinamento da MLP para a previsão do **absenteísmo trimestral** medido em horas.

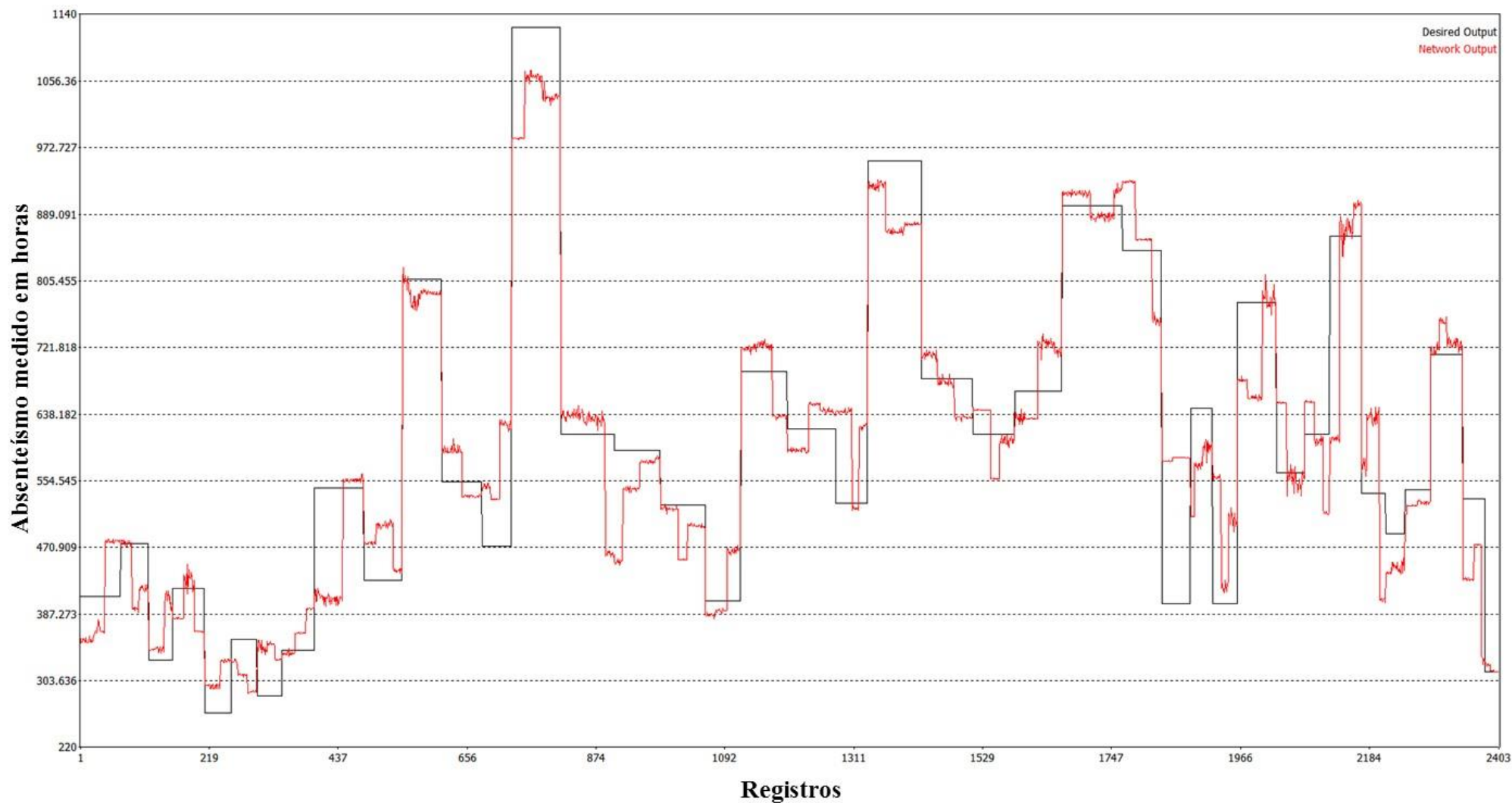


Figura 30 – Fase de treinamento da MLP para a previsão do absenteísmo **trimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245. Percebe-se que nos intervalos de 437 a 656, 874 a 1092 e de 1747 a 1966 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 31 mostra a fase de treinamento da MLP para a previsão do **absenteísmo quadrimestral** medido em horas.

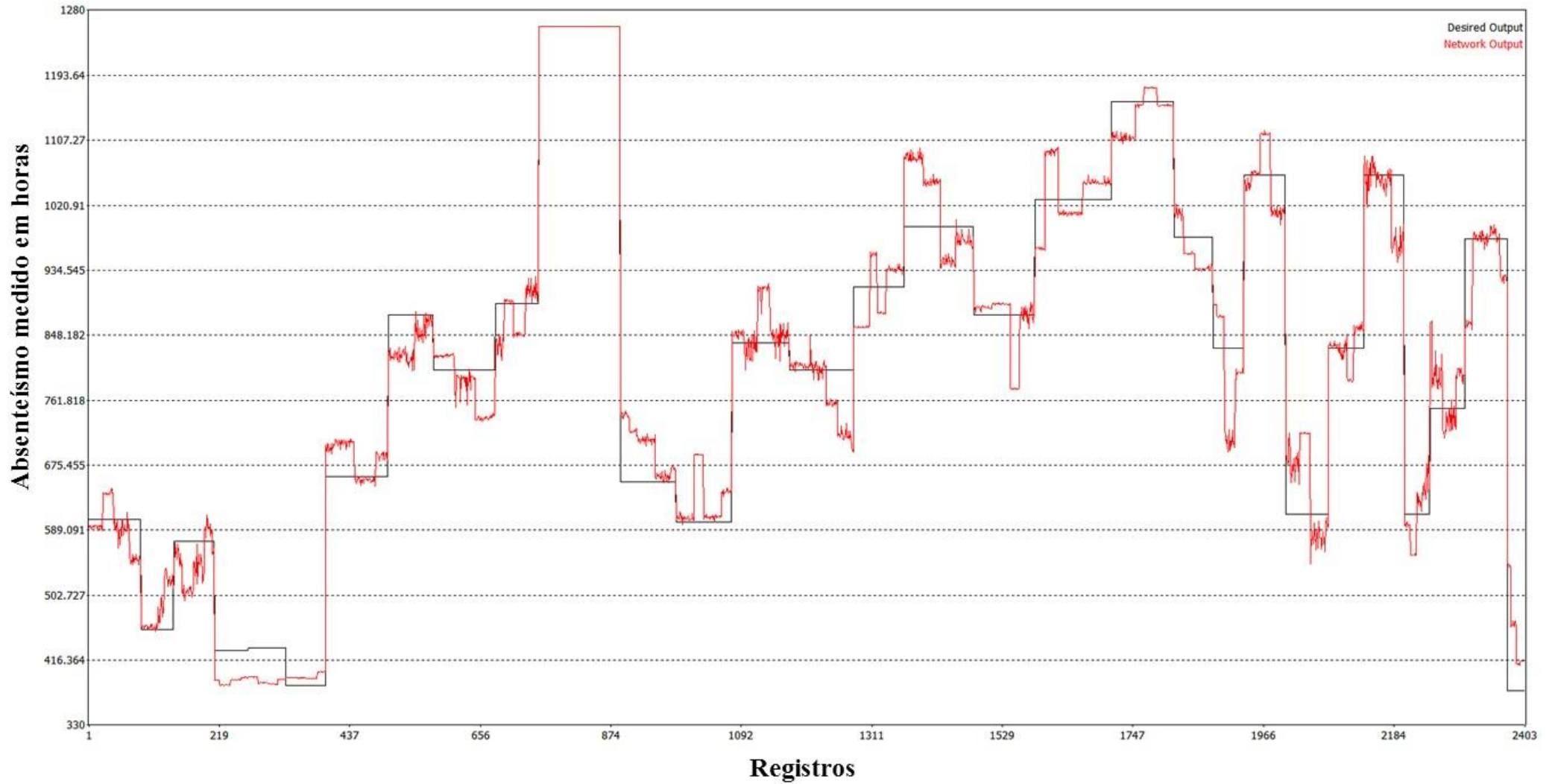


Figura 31 – Fase de treinamento da MLP para a previsão do absenteísmo **quadrimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245. Percebe-se que nos intervalos de 219 a 437 e de 1311 a 1529 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 32 mostra a fase de treinamento da MLP para a previsão do **absenteísmo semestral** medido em horas.

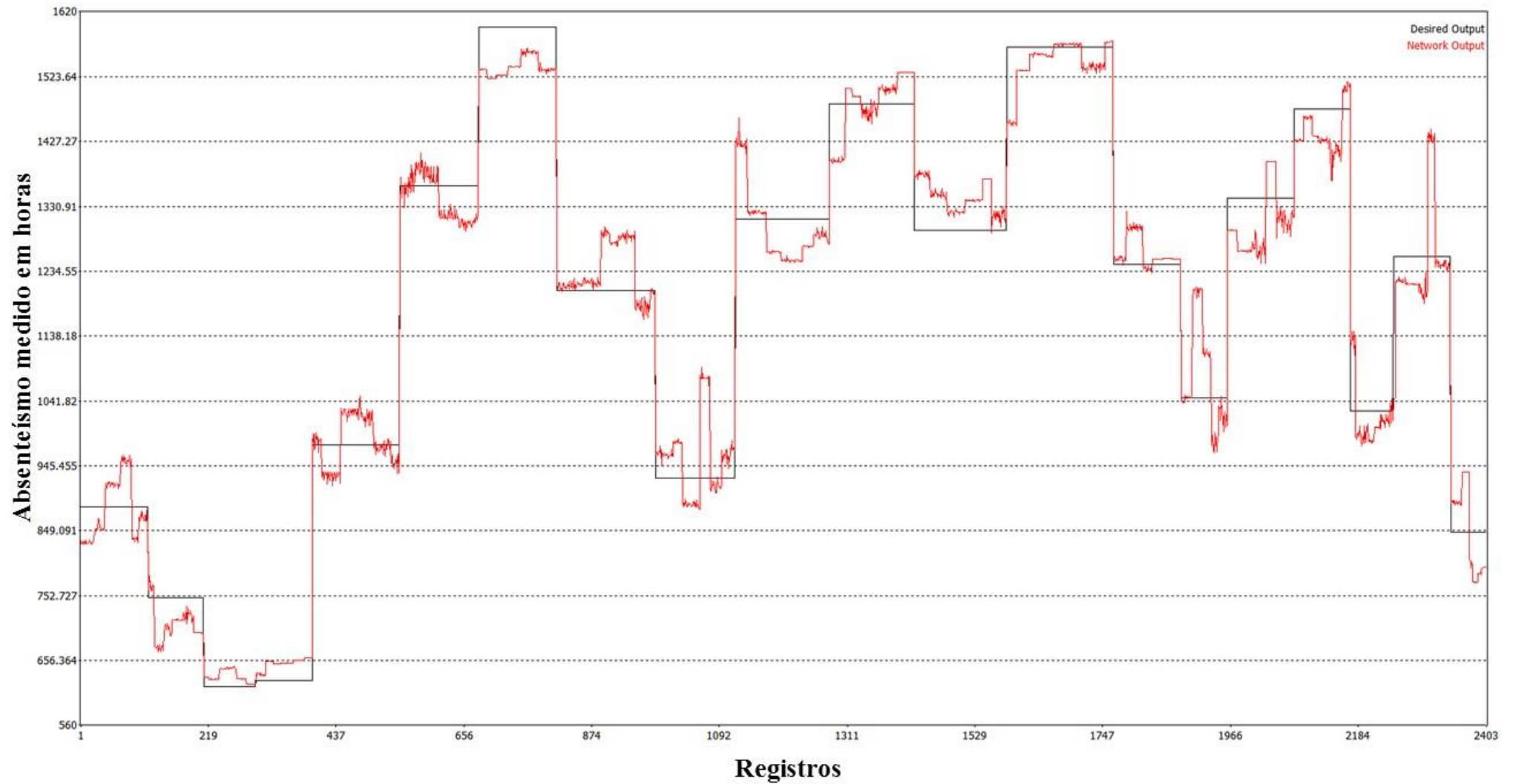


Figura 32 – Fase de treinamento da MLP para a previsão do absenteísmo **semestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245. Percebe-se que no intervalo de 2184 a 2403 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 33 mostra a fase de treinamento da MLP para a previsão do absenteísmo **anual** medido em horas.

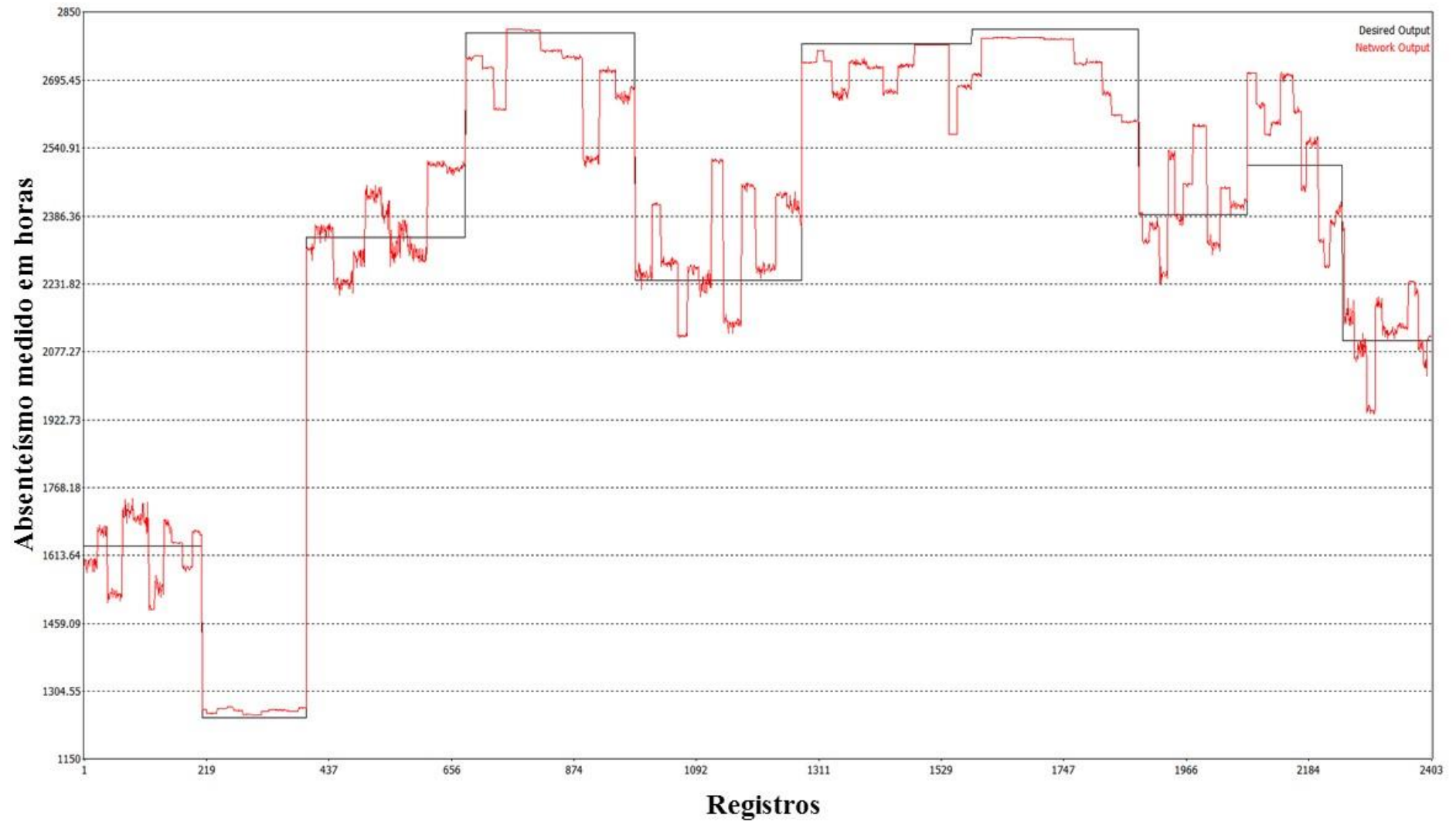


Figura 33 – Fase de treinamento da MLP para a previsão do absenteísmo **anual** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245. Percebe-se que nos intervalos de 874 a 1311 e de 1966 a 2184 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 34 mostra a fase de treinamento da MLP para a previsão do absentismo **bianual** medido em horas.

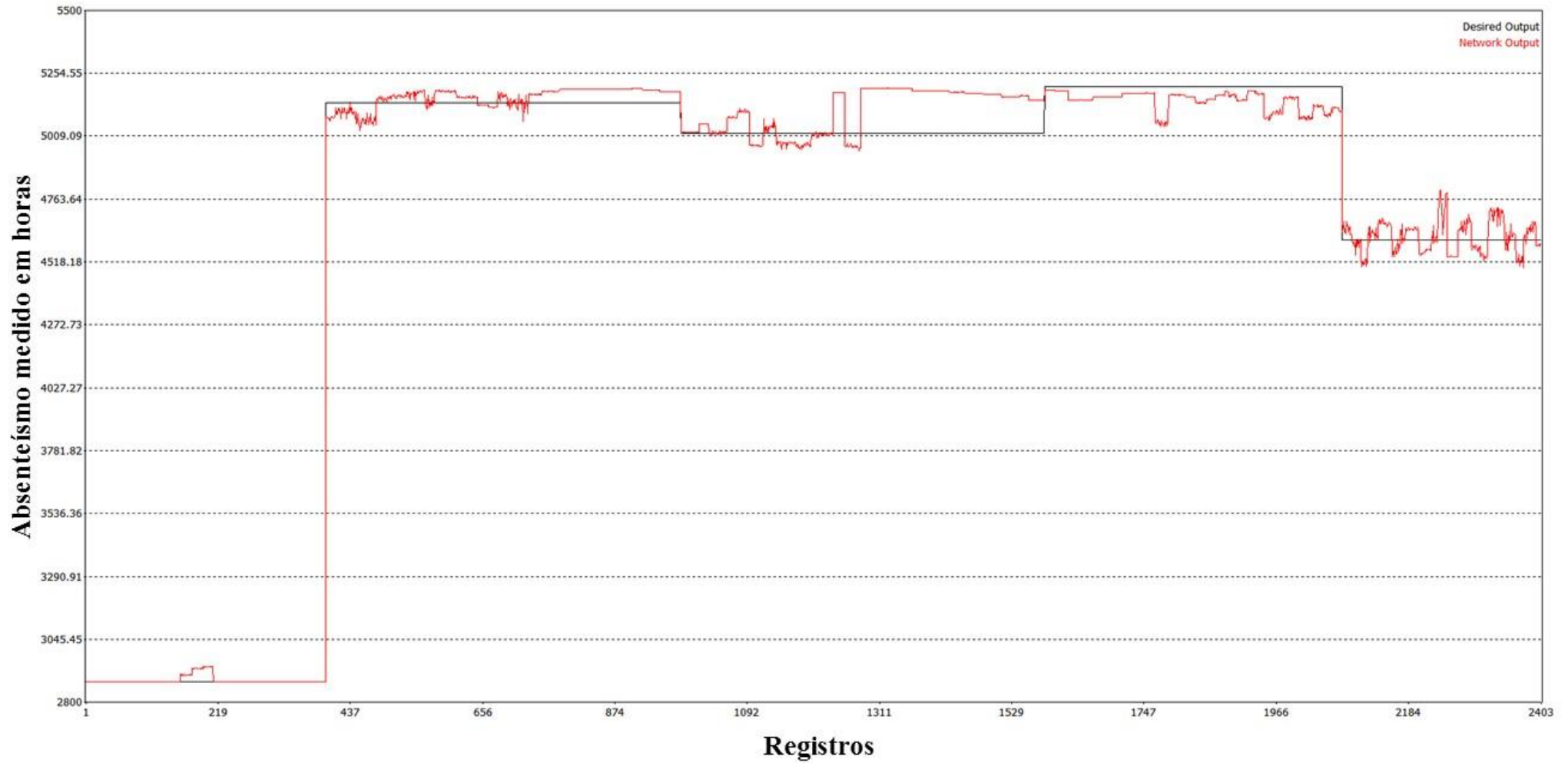


Figura 34 – Fase de treinamento da MLP para a previsão do absenteísmo **bianual** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0346047245. Exceto o final do terceiro biênio, intervalo entre 1311 a 1529 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada e o longo período para calcular a estimativa.

A Figura 35 mostra a curva do erro quadrático médio na fase de treinamento. O eixo x possui o intervalo de 0 a 8.500 e corresponde ao número de épocas de treinamento. O erro quadrático médio na fase de treinamento foi igual a 0,0346047245.

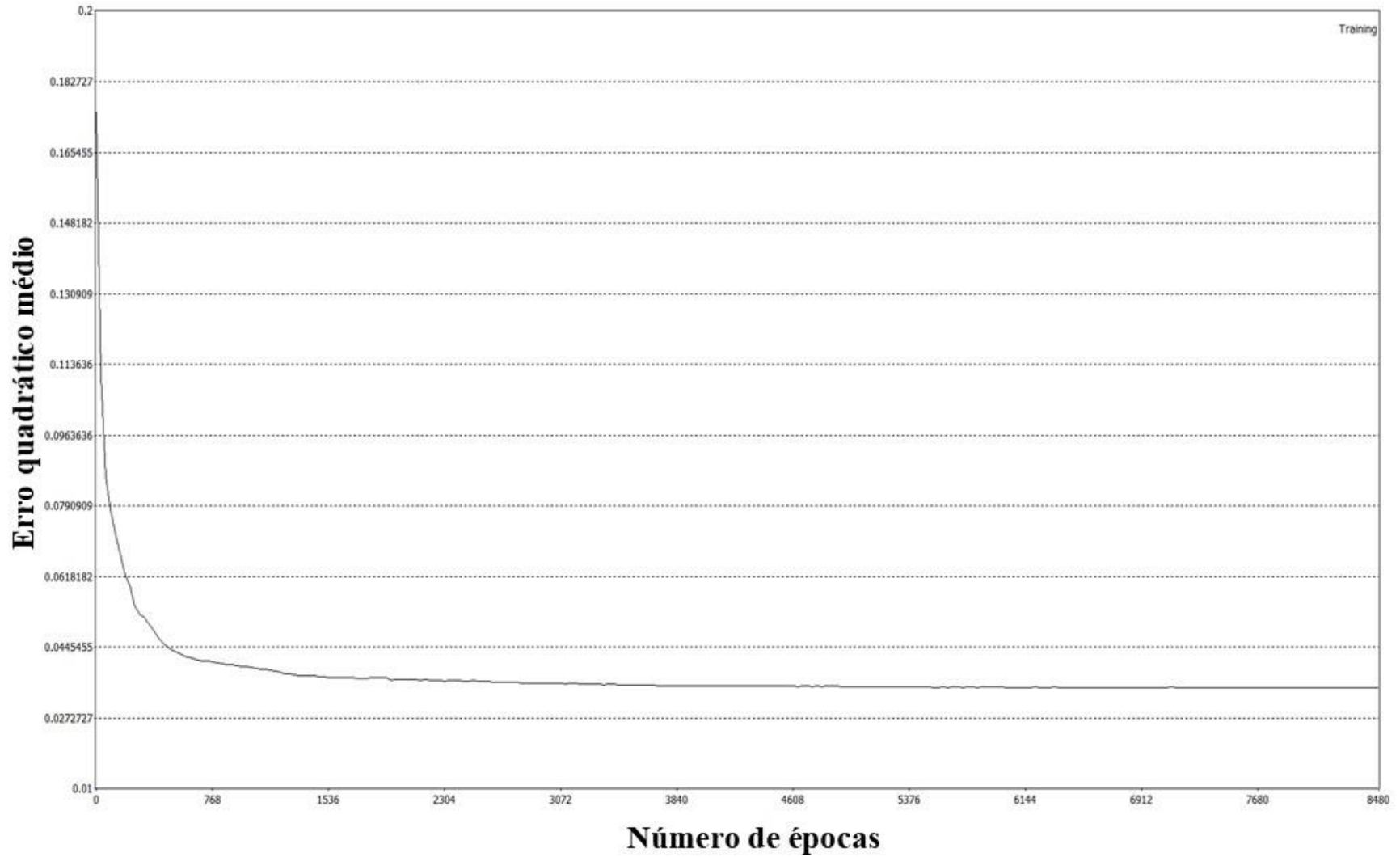


Figura 35 – Curva do erro quadrático médio na fase de **treinamento**.

Observa-se que a MLP na fase de treinamento apresentou o erro quadrático médio pequeno, o que indica capacidade de generalizar os dados de saída desejados. Percebe-se que não ocorreu oscilação durante a fase treinamento.

Os resultados apresentados na previsão do absentismo medido em horas na fase de treinamento mostraram que a técnica conseguiu representar os dados de saída desejados nos períodos propostos (mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual).

4.3 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos *Rough Sets* e MLP na previsão do absentismo (Experimento 1 - Treinamento).

Na Etapa 2, Experimento 1, a base de dados foi reduzida com *Rough Sets*. A base de dados inicialmente com 50 atributos foi reduzida utilizando **algoritmo genético** gerando 144 redutos.

A Tabela 18 mostra a distribuição dos 144 redutos gerados e o número de atributos com a base reduzida.

Tabela 18 – Redutos gerados utilizando **algoritmo genético** e a base reduzida

Atributos	Redutos Gerados	Base Reduzida
50	2	20
50	11	21
50	15	22
50	18	23
50	37	24
50	61	25

Optou-se pelo experimento com 20 atributos, ou seja, maior redução de atributos 60%.

A Figura 36 ilustra a realização dos experimentos da Fase 1, Etapa 2, Experimento 1 - Treinamento.

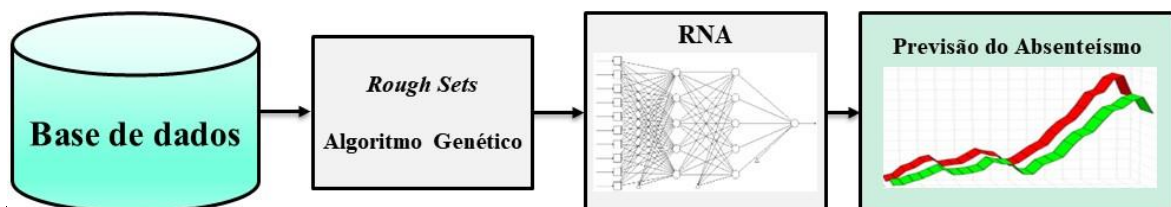


Figura 36 – Realização dos experimentos da Fase 1, Etapa 2, Experimento 1 - Treinamento.

Dos 20 atributos da base de dados reduzida 17 serviram de entradas da MLP e 3 atributos de saídas: absentismo medido em horas mensal, trimestral e quadrimestral. Os experimentos da Fase 1, Etapa 2, Experimento 1 foram realizados com o objetivo de prever o absentismo

mensal, trimestral e quadrimestral. A Tabela 19 apresenta os 17 atributos de entrada da MLP (Experimento 1).

Tabela 19 – Atributos de entrada da MLP (17 atributos)

Item	Atributos	Mínimo	Máximo
1	CID / Motivo da ausência	1	27
2	Dia da Semana	2	6
3	Estação do ano	1	4
4	Tempo de afastamento em dias	0	15
5	Tempo de afastamento em horas	0	7
6	Tempo de afastamento (dia+horas) em horas	1	120
7	Quantidade de atestados por mês	4	48
8	Valor do transporte para o trabalho mensal em reais (R\$)	118,8	388,7
9	Tempo de serviço (Em anos)	0	57
10	Tempo de serviço (Em meses)	6	695
11	Tempo de serviço (Em dias)	196	21160
12	Idade (Em dias)	9039	26882
13	Carga de trabalho/mês	4236560	10350518
14	Carga de trabalho (Média/dia)	192571	470478
15	Meta atingida	81,4	100
16	Despesas Médico-hospitalares e odontológica por idade (%)	2,5	47,4
17	Frequência Atestado Médico	0	1196

A Tabela 20 apresenta os parâmetros utilizados na MLP com 17 neurônios de entrada.

Tabela 20 – Parâmetros utilizados na MLP (17 neurônios de entrada)

Parâmetros	Valor
Número de neurônios de entrada	17
Número de camadas ocultas	2
Número de neurônios nas camadas ocultas	10
Taxa inicial de aprendizado	0,7
Decaimento a cada 500 épocas	1%
Fator de momento inicial	0,7
Decaimento a cada 1000 épocas	1%

A função de ativação dos neurônios das camadas ocultas e das saídas da MLP foi a sigmoide, os pesos foram inicializados aleatoriamente entre -1 e 1, o *bias* fixado em -1, o critério de parada foi o erro quadrático médio na fase de treinamento menor que 0,01 ou 8.500 épocas.

A Figura 37 ilustra a topologia da MLP utilizada no experimento Fase 1, Etapa 2, Experimento 1 - Treinamento.

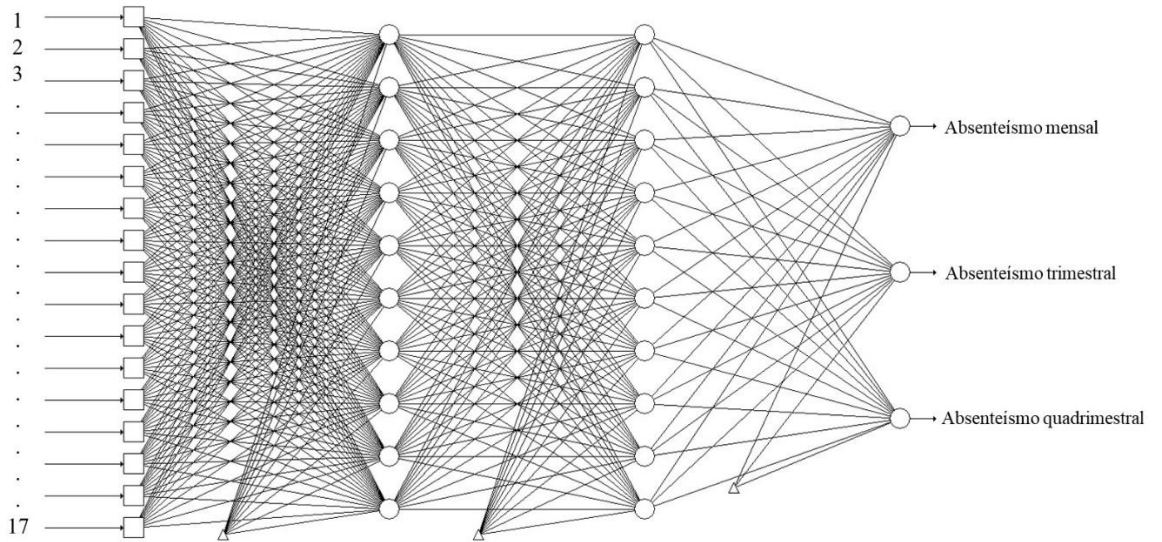


Figura 37 – Topologia da MLP utilizada no Experimento 1.

O treinamento da MLP foi sequencial com 8.500 épocas e o tempo de processamento foi de 9 minutos e 15 segundos. O erro quadrático médio na fase de treinamento foi igual a 0,0516407691. As saídas da MLP foram o absenteísmo medido em horas mensal, trimestral e quadrimestral.

A Figura 38 mostra a fase de treinamento da MLP para a previsão do absenteísmo **mensal** medido em horas. A curva na cor preta é a saída desejada e a curva na cor vermelha a saída da MLP.

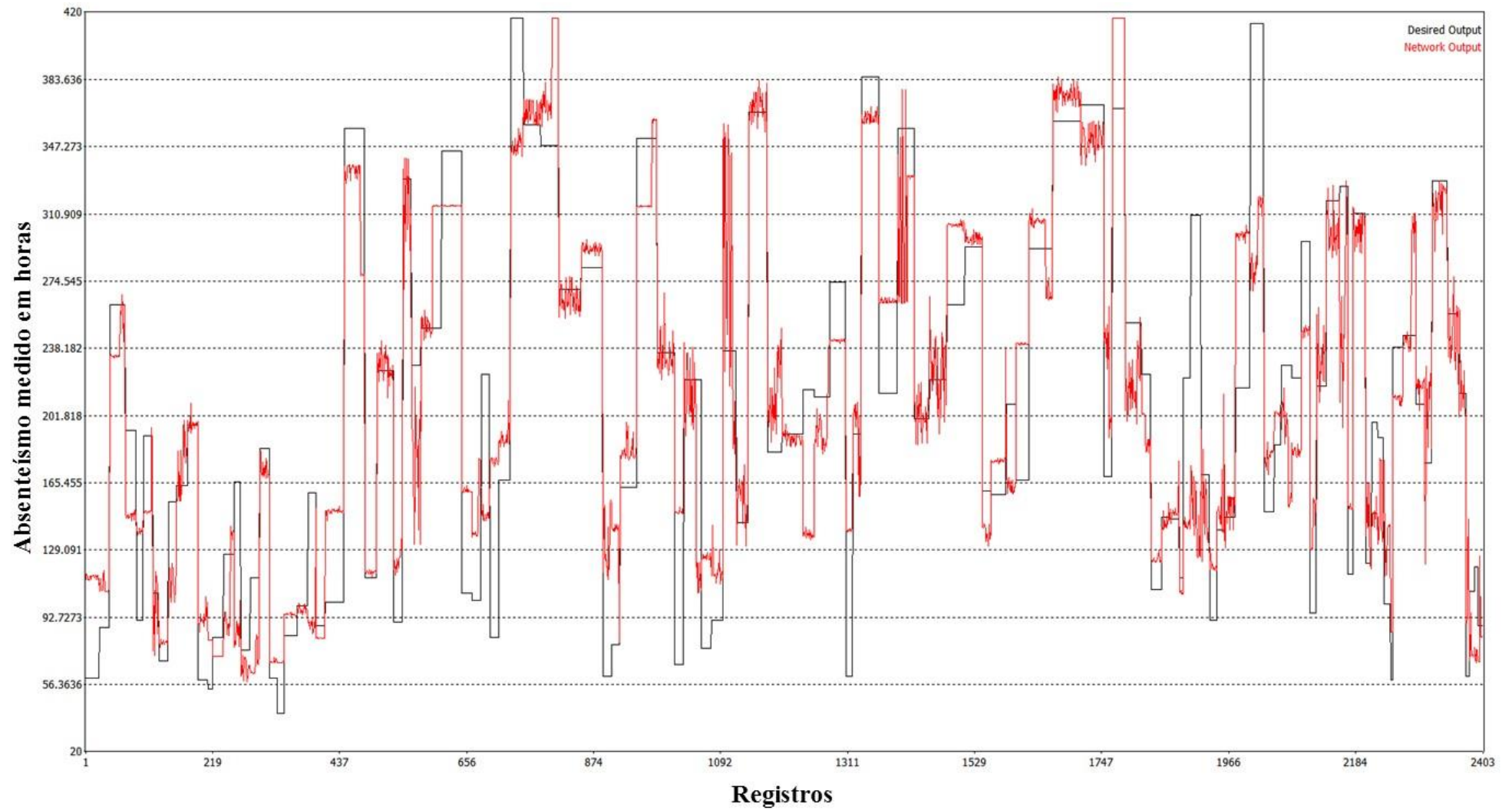


Figura 38 – Fase de treinamento da MLP para a previsão do absenteísmo **mensal** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0516407691. Percebe-se que nos intervalos de 656 a 1092 e de 1747 a 2184 referentes aos registros as estimativas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 39 mostra a fase de treinamento da MLP para a previsão do **absenteísmo trimestral** medido em horas.

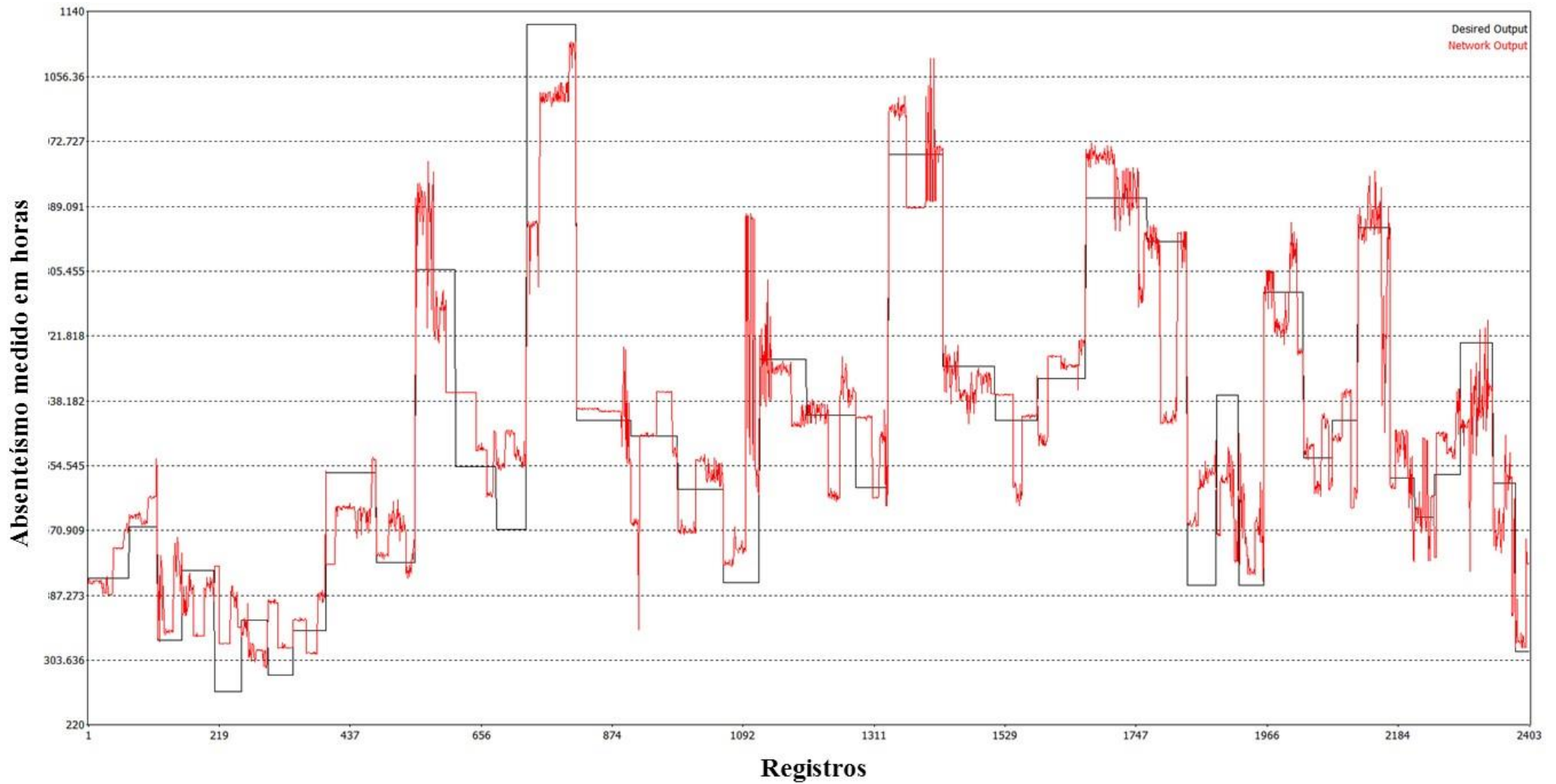


Figura 39 – Fase de treinamento da MLP para a previsão do absenteísmo **trimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0516407691. Percebe-se que nos intervalos de 656 a 874, 1092 a 1311 e de 1747 a 1966 referentes aos registros as estimativas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 40 mostra a fase de treinamento da MLP para a previsão do **absenteísmo quadrimestral** medido em horas.

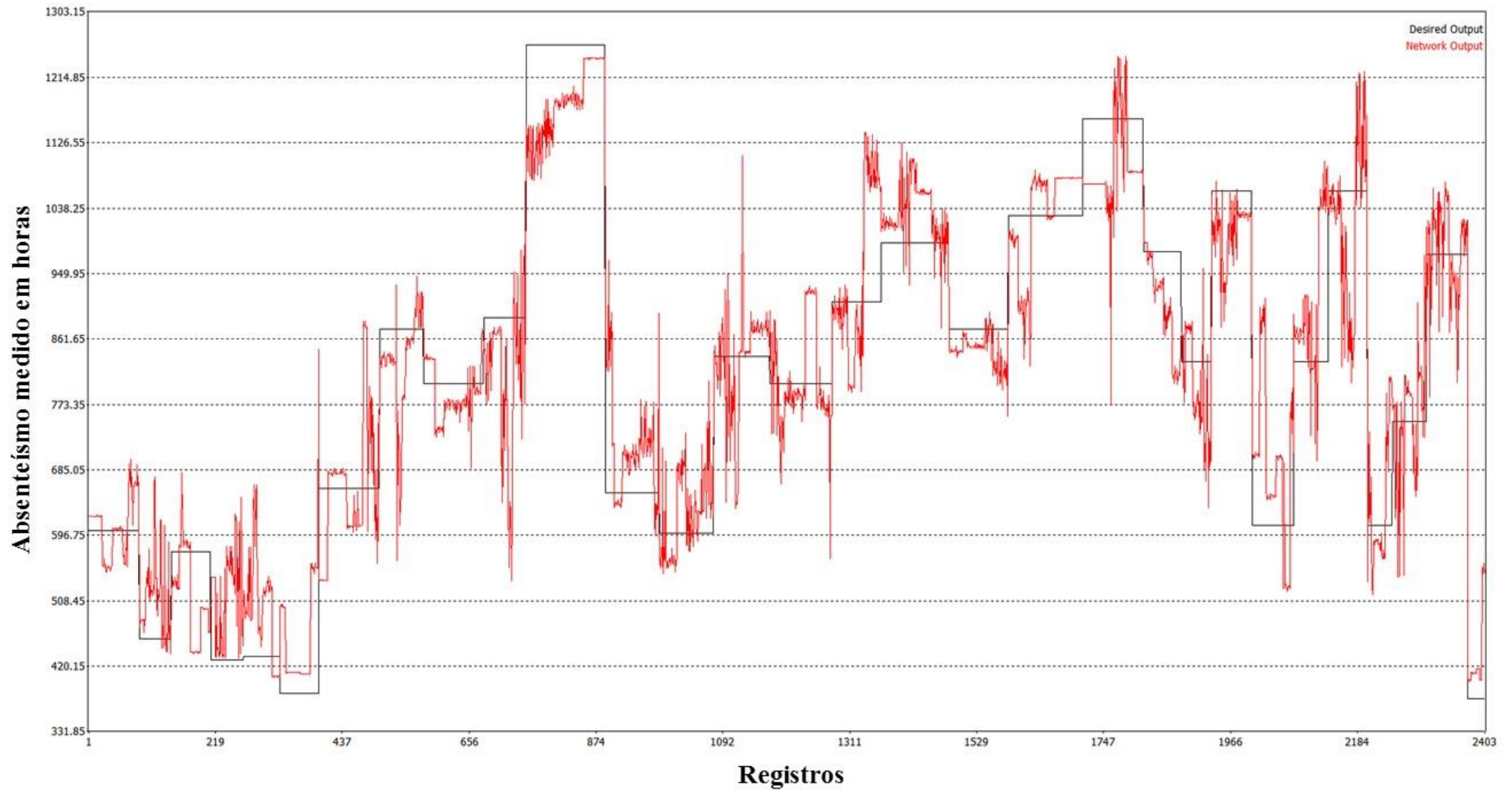


Figura 40 – Fase de treinamento da MLP para a previsão do absenteísmo **quadrimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0516407691. Percebe-se que nos intervalos de 656 a 1311 e de 1747 a 2184 referentes aos registros as estimativas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 41 mostra a curva do erro quadrático médio na fase de treinamento. O eixo x possui o intervalo de 0 a 8.500 e corresponde ao número de épocas de treinamento. O erro quadrático médio na fase de treinamento foi igual a 0,0516407691. Percebe-se que ocorreu uma pequena oscilação durante toda a fase treinamento, provavelmente devido aos ruídos característicos da base de dados utilizada.

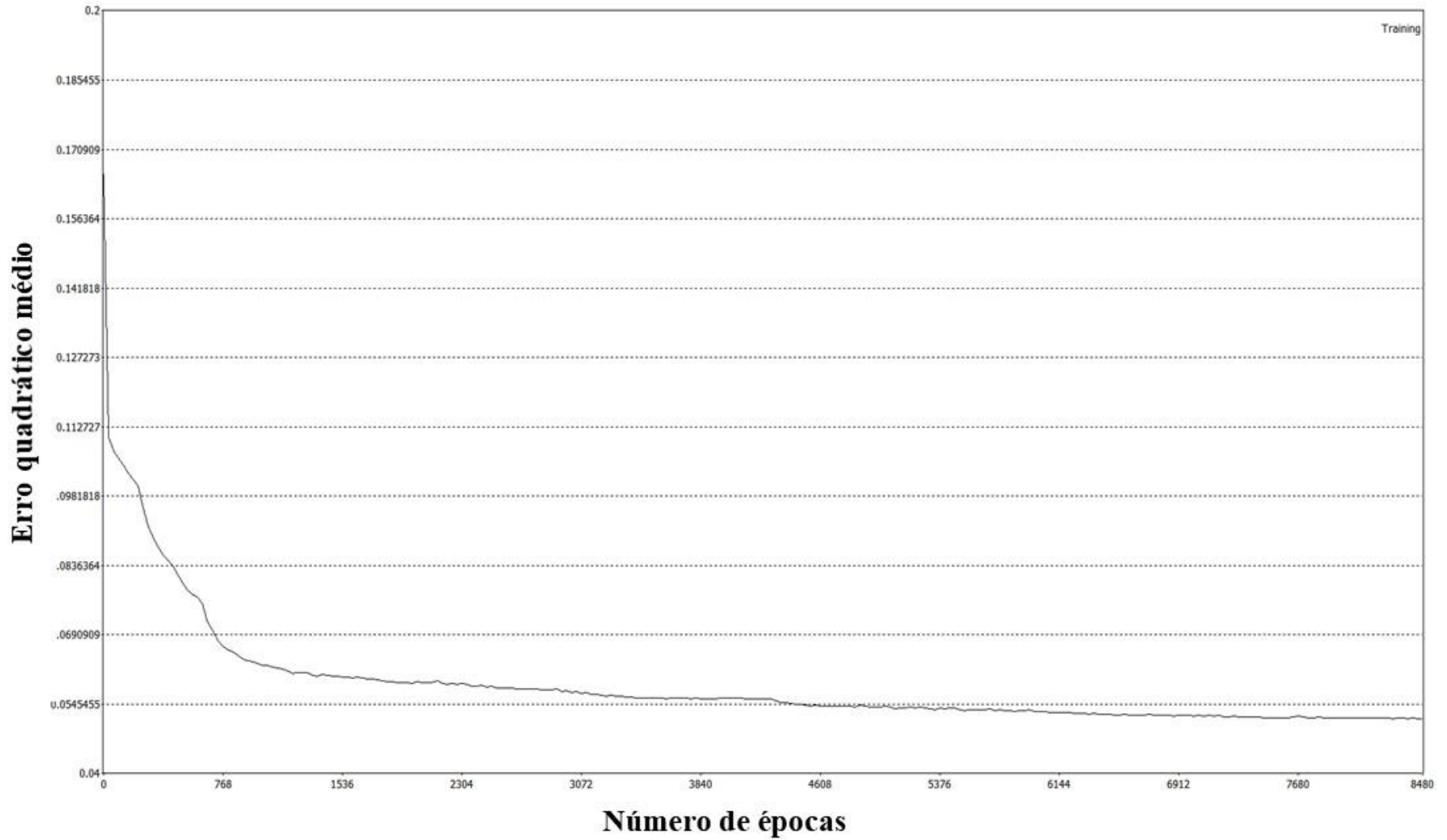


Figura 41 – Curva do erro quadrático médio na fase de **treinamento**.

Observa-se no experimento com a MLP que os atributos reduzidos por *Rough Sets* com o método de redução por algoritmo genético apresentaram o erro quadrático médio igual a 0,0516407691, o que indica capacidade de generalizar os dados de saída desejados. Percebe-se que ocorreu uma pequena oscilação durante a fase treinamento, provavelmente essa oscilação seja causada devido aos ruídos característicos da base de dados utilizada e também pela redução dos atributos.

Os resultados apresentados na previsão do absenteísmo com o *Rough Sets* e a MLP na fase de treinamento mostraram que as duas técnicas conseguiram representar os dados de saída desejados nos períodos propostos mensal, trimestral e quadrimestral apesar da redução dos atributos.

O treinamento realizado com a redução de 60% dos atributos por *Rough Sets* foi capaz de realizar a previsão mensal, trimestral e quadrimestral mesmo aumentando o erro quadrático médio em 0,0170360446 em relação ao experimento com a base de dados completa (50 atributos).

Dessa forma, entende-se que foi possível realizar a previsão do absenteísmo com a redução de 50 atributos para 20 atributos o que viabiliza sua utilização em situações em que a empresa por diferentes motivos não possa coletar os cinquenta atributos como apresentado nos experimentos da Etapa 1.

4.4 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos *Rough Sets* e MLP na previsão do absenteísmo (Experimento 2 - Treinamento).

Na Etapa 2 (Experimento 2 - Treinamento) a base de dados foi reduzida com *Rough Sets*, a base de dados inicialmente com 50 atributos foi reduzida utilizando o método de redução pelo algoritmo de Johnson gerando um único reduto com 19 atributos, ou seja, redução de 62% dos atributos.

O Experimento 2 foi realizado com o objetivo de prever o absenteísmo mensal, trimestral e quadrimestral.

A Figura 42 ilustra a realização dos experimentos da Etapa 2 (Experimento 2 - Treinamento).

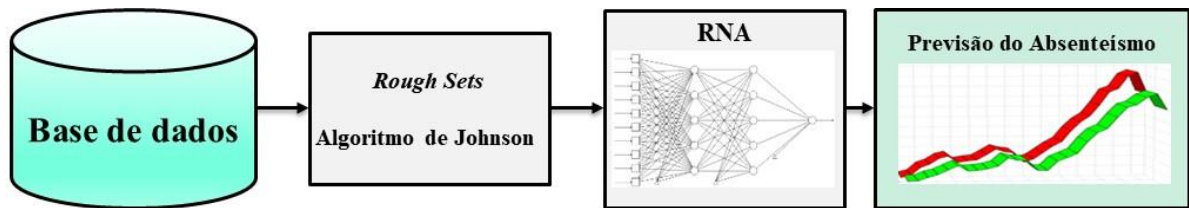


Figura 42 – Realização dos experimentos da Etapa 2 (Experimento 2 - Treinamento).

Dos 19 atributos da base de dados reduzida 16 serviram de entradas da MLP e 3 atributos de saídas: absenteísmo medido em horas mensal, trimestral e quadrimestral. Os experimentos da Fase 1, Etapa 2 (Experimento 2) foram realizados com o objetivo de prever o absenteísmo mensal, trimestral e quadrimestral. A Tabela 21 apresenta os 16 atributos de entrada da MLP (Experimento 2).

Tabela 21 – Atributos de entrada da MLP (16 atributos)

Item	Atributos	Mínimo	Máximo
1	CID / Motivo da ausência	1	27
2	Dia da Semana	2	6
3	Estação do ano	1	4
4	Tempo de afastamento em dias	0	15
5	Tempo de afastamento em horas	0	7
6	Tempo de afastamento (dia+horas) em horas	1	120
7	Quantidade de atestados por mês	4	48
8	Valor do transporte para o trabalho mensal em reais (R\$)	118,8	388,7
9	Tempo de serviço (Em anos)	0	57
10	Tempo de serviço (Em meses)	6	695
11	Tempo de serviço (Em dias)	196	21160
12	Idade (Em dias)	9039	26882
13	Carga de trabalho/mês	4236560	10350518
14	Meta atingida	81,4	100
15	Despesas Médico-hospitalares e odontológica por idade (%)	2,5	47,4
16	Frequência Atestado Médico	0	1196

A Tabela 22 apresenta os parâmetros utilizados na MLP com 16 neurônios de entrada.

Tabela 22 – Parâmetros utilizados na MLP (16 neurônios de entrada)

Parâmetros	Valor
Número de neurônios de entrada	16
Número de camadas ocultas	2
Número de neurônios nas camadas ocultas	10
Taxa inicial de aprendizado	0,7
Decaimento a cada 500 épocas	1%
Fator de momento inicial	0,7
Decaimento a cada 1000 épocas	1%

A função de ativação dos neurônios das camadas ocultas e das saídas da RNA foi a sigmoide, os pesos foram inicializados aleatoriamente entre -1 e 1, o *bias* fixado em -1, o critério de parada foi o erro quadrático médio na fase de treinamento menor que 0,01 ou 8.500 épocas. A Figura 43 ilustra a topologia utilizada na Fase 1, Etapa 2 (Experimento 2 - Treinamento).

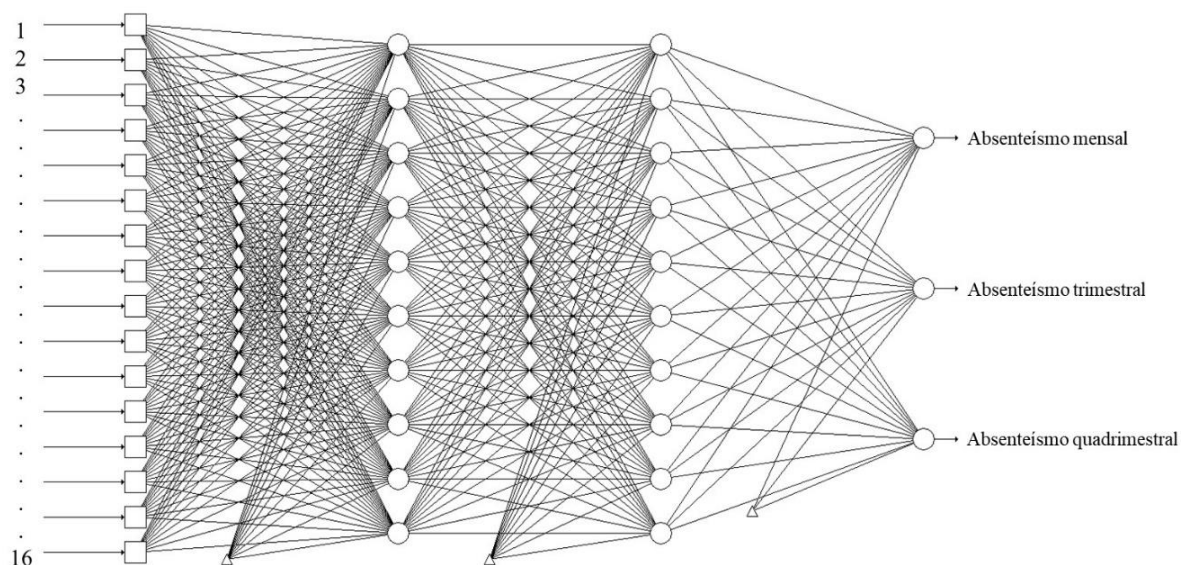


Figura 43 – Topologia da MLP utilizada na Fase 1, Etapa 2 (Experimento 2 - Treinamento).

O treinamento da MLP foi sequencial com 8.500 épocas e o tempo de processamento foi de 8 minutos e 44 segundos. O erro quadrático médio na fase de treinamento foi igual a

0,0575417573. As saídas foram o absenteísmo medido em horas mensal, trimestral e quadrimestral.

A Figura 44 mostra a fase de treinamento para a previsão do absenteísmo **mensal** medido em horas. A curva na cor preta é a saída desejada e a curva na cor vermelha a saída da MLP.

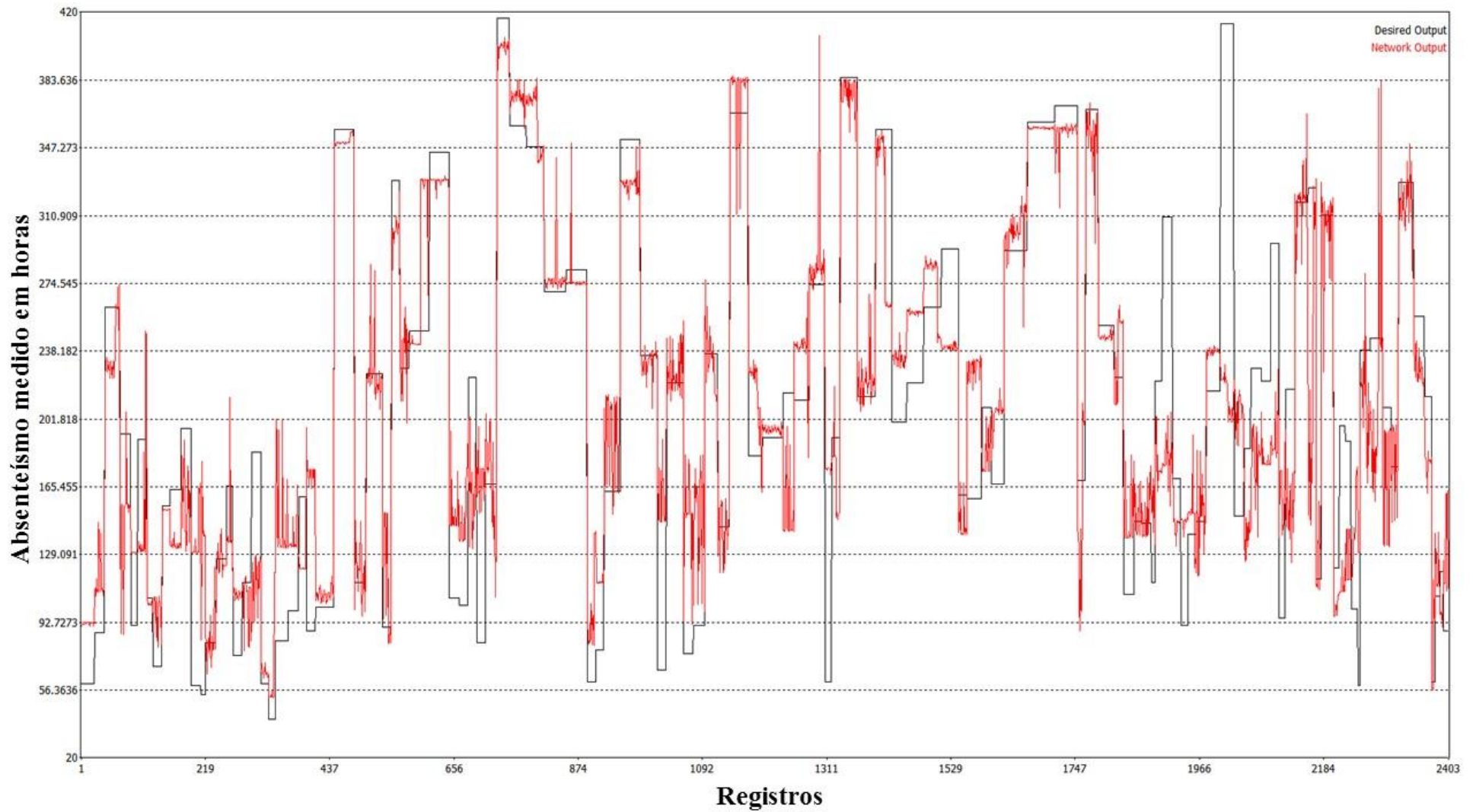


Figura 44 – Fase de treinamento da MLP para a previsão do absenteísmo **mensal** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0575417573. Percebe-se que nos intervalos de 874 a 1529 e de 1747 a 2184 referentes aos registros as estimativas calculadas pela MLP apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 45 mostra a fase de treinamento da MLP para a previsão do **absenteísmo trimestral** medido em horas.

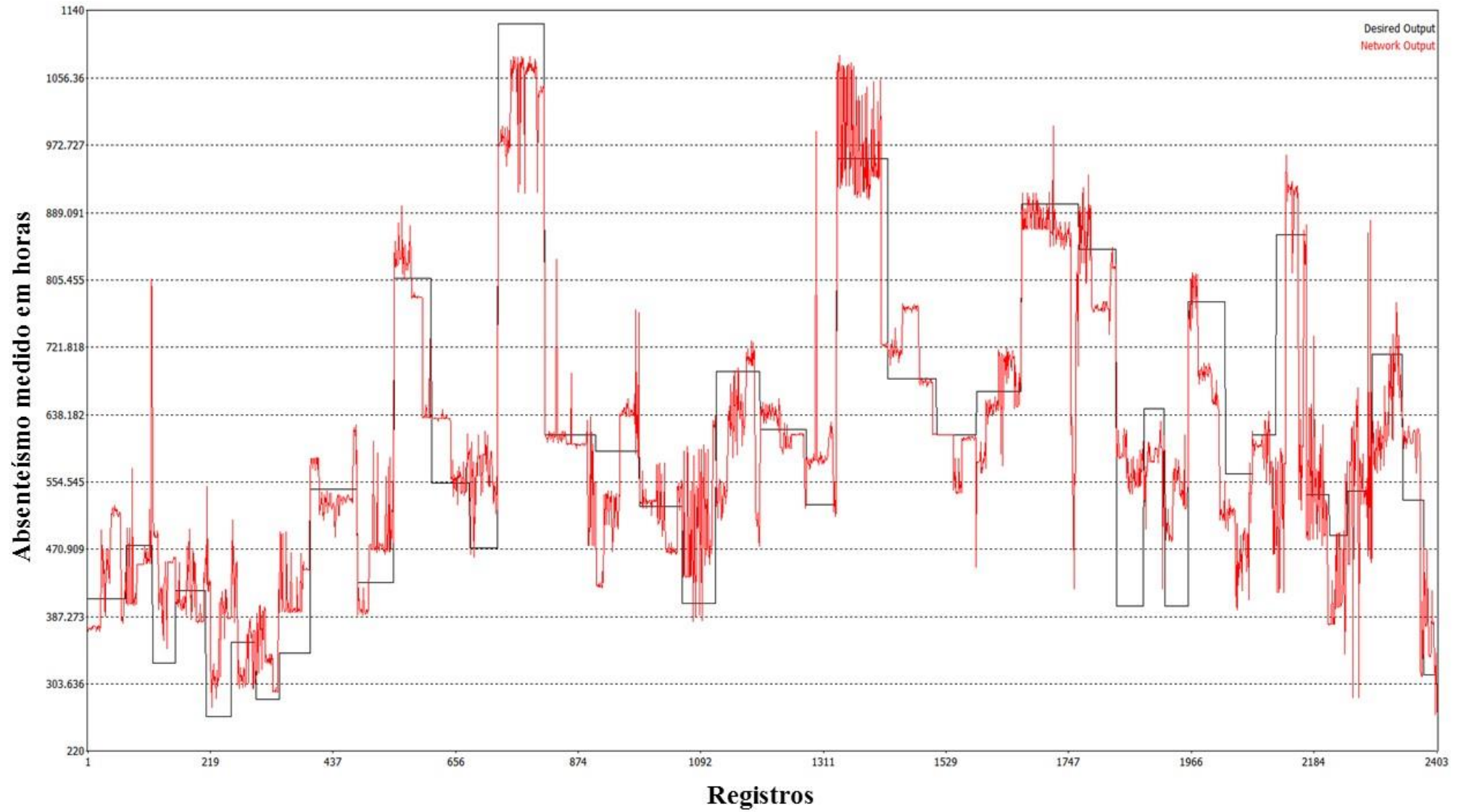


Figura 45 – Fase de treinamento da MLP para a previsão do absenteísmo **trimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0575417573. Percebe-se que nos intervalos de 656 a 874 e de 1747 a 2403 referentes aos registros as estimativas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 46 mostra a fase de treinamento da MLP para a previsão do **absenteísmo quadrimestral** medido em horas.

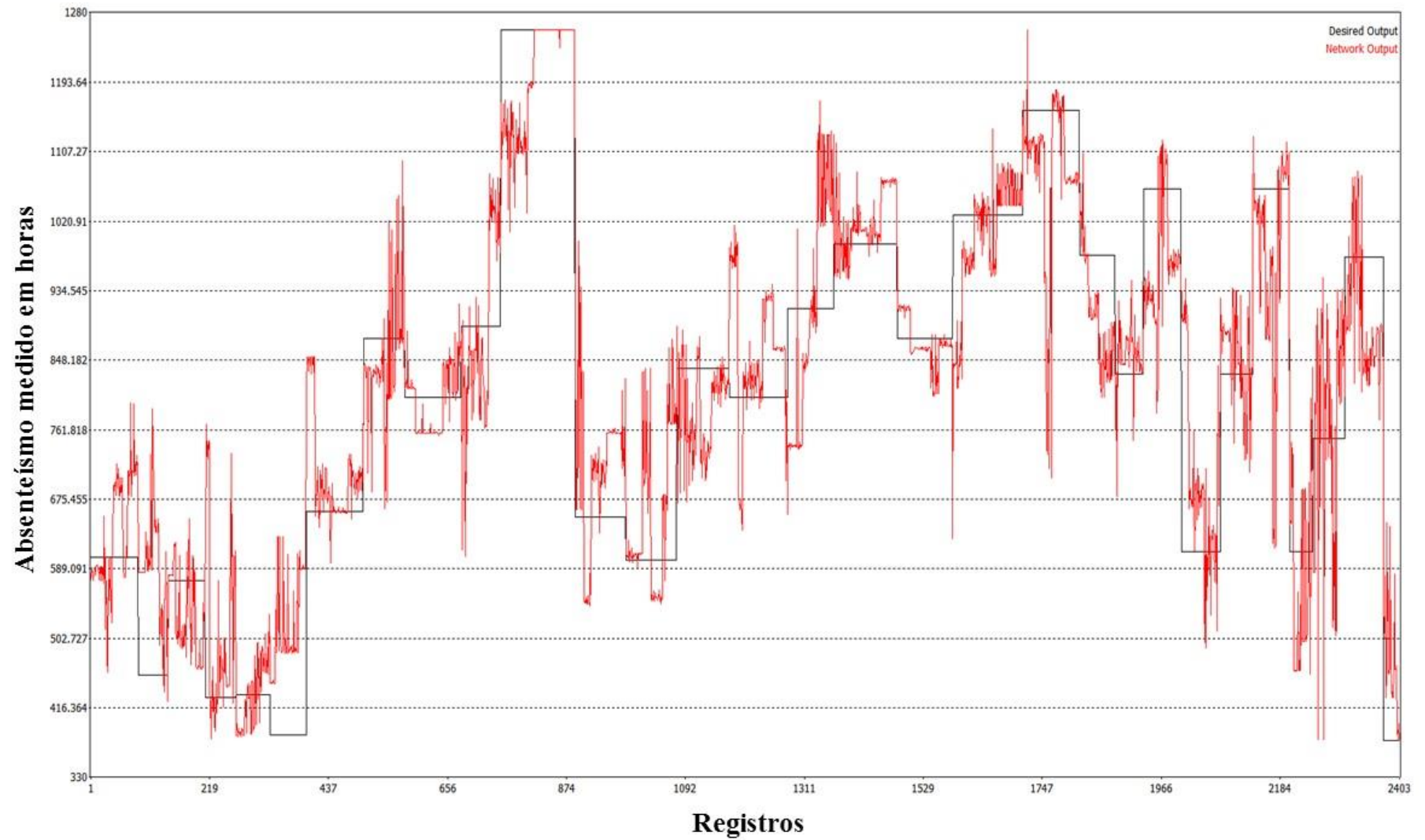


Figura 46 – Fase de treinamento da MLP para a previsão do absentismo **quadrimestral** medido em horas.

Observa-se que a MLP na fase de treinamento conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0575417573. Percebe-se que nos intervalos de 1 a 874, 1529 a 1966 e de 2184 a 2403 referentes aos registros as estimativas calculadas pela MLP apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Figura 47 mostra a curva do erro quadrático médio na fase de treinamento da MLP. O eixo x possui o intervalo de 0 a 8.500 e corresponde as épocas de treinamento. Percebe-se que ocorreu uma pequena oscilação durante a fase treinamento no intervalo de 0 a 768 referentes aos registros, provavelmente essa oscilação seja causada devido aos ruídos característicos da base de dados utilizada e também pela redução dos atributos.

Os resultados apresentados na previsão do absenteísmo com o *Rough Sets* e a MLP na fase de treinamento mostraram que as duas técnicas reunidas conseguiram representar os dados de saída desejados nos períodos propostos mensal, trimestral e quadrimestral apesar da redução dos atributos.

O treinamento realizado com a redução de 62% dos atributos por *Rough Sets* foi capaz de realizar a previsão mensal, trimestral e quadrimestral aumentando o erro quadrático médio em 0,0229370328 em relação ao experimento com a base de dados completa (50 atributos).

Dessa forma, entende-se que foi possível realizar a previsão do absenteísmo com a redução de 50 atributos para 19 atributos o que viabiliza sua utilização em situações em que a empresa por diferentes motivos não possa utilizar os cinquenta atributos como apresentado nos experimentos da Etapa 1.

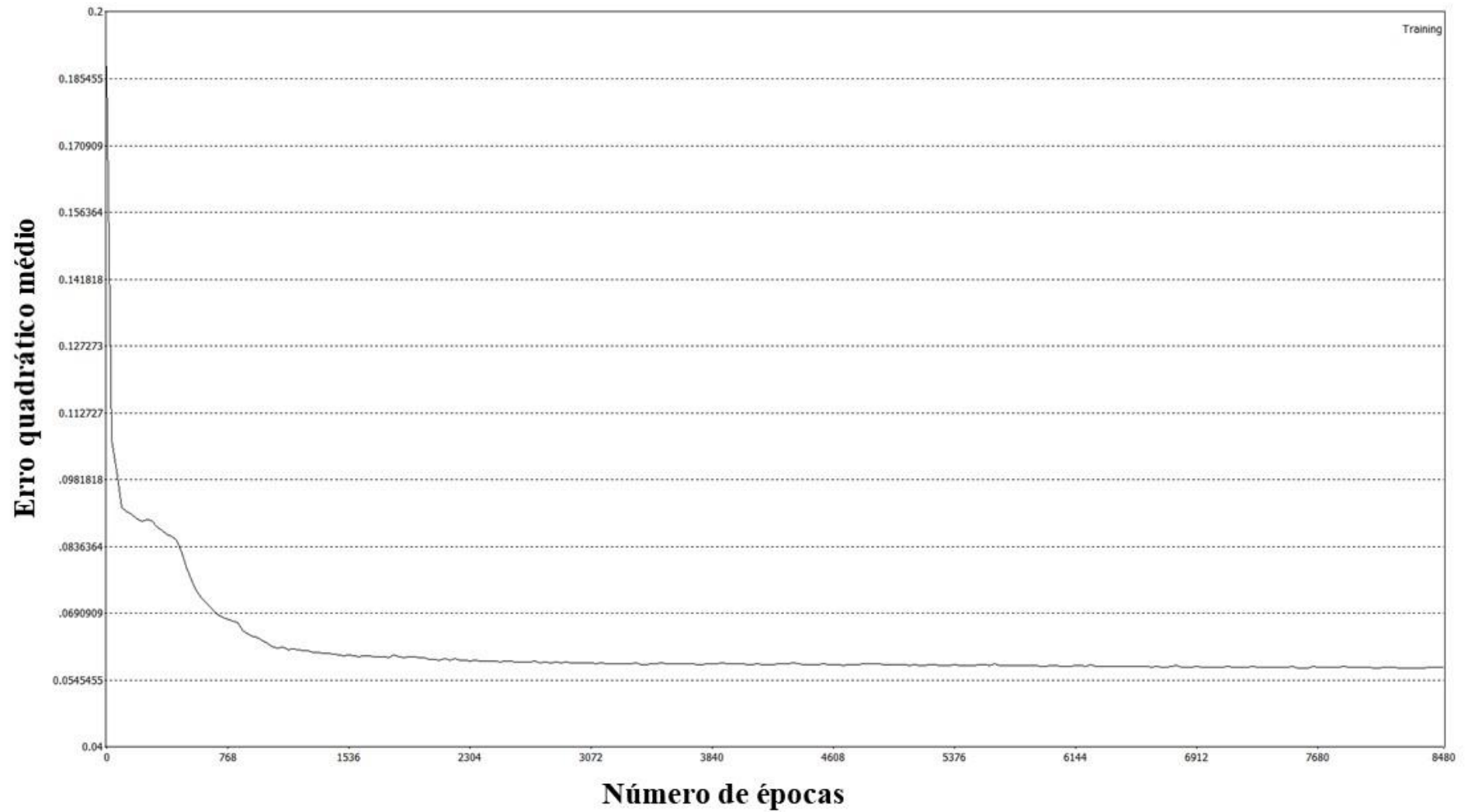


Figura 47 – Curva do erro quadrático médio na fase de **treinamento**.

Observa-se no experimento com a MLP com os atributos reduzidos pelos *Rough Sets* com o método de redução pelo algoritmo de Johnson apresentou o erro quadrático médio igual a 0,0575417573, o que indica capacidade de generalizar os dados de saída desejados.

Os resultados apresentados na previsão do absentismo com a MLP na fase de treinamento mostraram que a técnica conseguiu representar os dados de saída desejados nos períodos propostos mensal, trimestral e quadrimestral. O treinamento com a redução de 62% dos atributos por *Rough Sets* foi capaz de realizar a previsão mensal, trimestral e quadrimestral.

Na seção seguinte apresenta-se uma síntese dos experimentos computacionais da Fase 1 (Treinamento).

4.5 Síntese dos resultados dos experimentos computacionais Fase 1 (Treinamento)

O Quadro 5 apresenta a síntese dos resultados dos experimentos computacionais realizados na Fase 1, conforme descrição a seguir:

-Etapa 1 experimento com a RNA com a base de dados completa com 50 atributos;

-Etapa 2 Experimento 1 com a base de dados reduzida pelo método de redução por algoritmo genético e Experimento 2 com a base de dados reduzida pelo método de redução por algoritmo Johnson.

Quadro 5 – Síntese dos resultados dos experimentos computacionais da Fase 1 (Treinamento)

Experimentos Fase 1	Número de atributos	Número de entradas	Número de saídas	Tempo de processamento	Saídas Absenteísmo medido em horas	Erro quadrático médio	Diferença - Erro quadrático médio em relação ao experimento (Etapa 1 - 50 atributos)	Redução dos Atributos
Etapa 1	50	43	7	17 minutos e 13 segundos	Mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual	0,0346047245	-	-
Etapa 2 Experimento 1	20	17	3	9 minutos e 15 segundos	Mensal, trimestral e quadrimestral	0,0516407691	0,0170360446	-60%
Etapa 2 Experimento 2	19	16	3	8 minutos e 44 segundos	Mensal, trimestral e quadrimestral	0,0575417573	0,0229370328	-62%

Observa-se que o experimento da Etapa 1 (50 atributos) apresentou menor erro quadrático médio e maior com número de saídas, embora tenha apresentado o maior tempo de processamento 17 minutos e 13 segundos. Os demais experimentos com redução de atributos

apresentaram resultados com erros quadráticos médios um pouco maiores, menos saídas e menor tempo de processamento.

Verifica-se na Etapa 2 – Experimento 1 com a redução de 50 para 20 atributos, ou seja, redução de 60% dos atributos da base de dados, o erro quadrático médio aumentou 0,0170360446 em relação ao experimento com a base de dados completa na Etapa 1. O tempo de processamento foi de 9 minutos e 15 segundos.

Na Etapa 2 – Experimento 2 com a redução de 50 para 19 atributos, ou seja, redução de 62% dos atributos da base de dados apresentou erro quadrático médio maior 0,0229370328 em relação ao experimento com a base de dados completa na Etapa 1.

Dessa forma, percebeu-se que os atributos desconsiderados por *Rough Sets* não representavam as características essenciais da base de dados. O tempo de processamento foi o menor 8 minutos e 44 segundos.

Na Etapa 2, os experimentos 1 e 2 com reduções de atributos apresentaram erro quadrático médio com pequeno aumento em relação ao experimento com a base de dados sem redução, no entanto, o tempo de processamento foi menor.

Nas etapas 1 e 2 foram apresentados os experimentos nas fases de treinamento em que se observa aderência.

A seguir são apresentados os experimentos computacionais da Fase 1, Etapas 1 e 2 na fase de teste.

4.6 Fase 1 - Etapa 1 (RNA): Aplicação de uma MLP na previsão do absenteísmo (Teste)

Para a fase de teste a base de dados com 2403 registros foi dividida em duas partes:

- A primeira para treinamento com 2.074 registros, ou seja, 86,3% dos registros equivalentes ao período de janeiro de 2008 a dezembro de 2015, considerando oito anos completos, 80% do período;

- A segunda para teste com 329 registros, ou seja, 13,7% dos registros equivalentes ao período de janeiro de 2016 a dezembro de 2017, considerando dois anos completos, 20% do período.

Optou-se pela divisão por período anual e não pela quantidade de dados de modo que os doze meses do ano fossem incluídos de forma igual na base de dados de treinamento e teste.

Vale destacar que foram testadas duas divisões da base de dados: seis anos para treinamento e quatro anos para teste e sete anos para treinamento e três anos para teste. Contudo,

a MLP não convergiu, por esse motivo, optou-se pela divisão em oito anos para treinamento e dois anos para teste.

Para medir o desempenho dos resultados dos experimentos foi comparada à saída desejada, absenteísmo medido em horas (real) e a estimativa calculada pela MLP. Dessa forma, foi calculado e apresentado o acerto da MLP na fase de teste.

A Figura 48 mostra a fase de teste da MLP, Fase 1, Etapa 1 para a previsão do absenteísmo mensal medido em horas. O eixo x apresenta o intervalo de 1 a 329 referentes aos registros que correspondem ao período de dois anos ou 13,7% dos registros. A curva na cor preta é a saída desejada e a curva na cor vermelha a saída da MLP. O tempo de processamento foi de 17 minutos e 48 segundos.

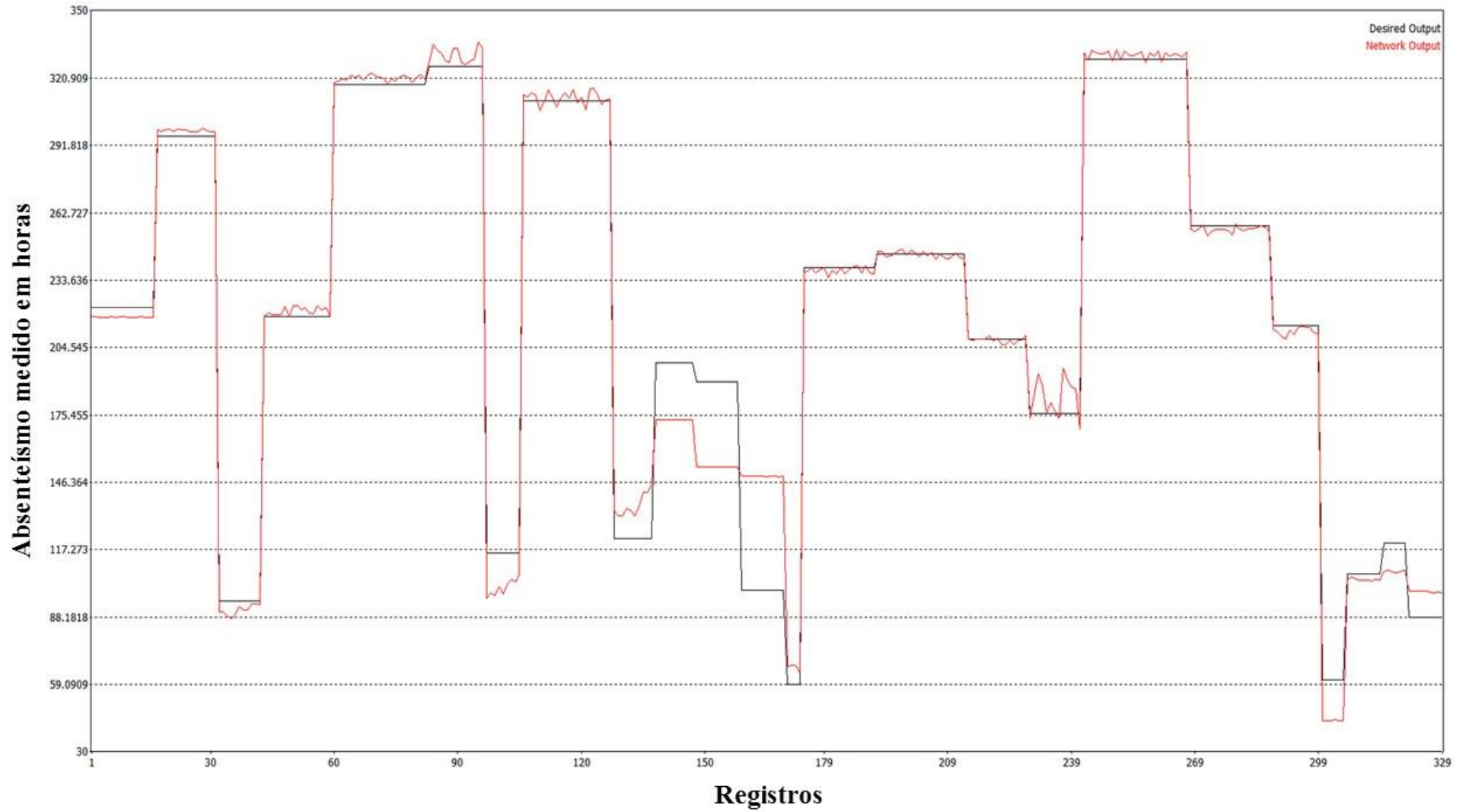


Figura 48 – Fase de teste da MLP para a previsão do absenteeismo **mensal** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0116904677 e na fase de teste igual a 0,0125162287. Percebe-se que no intervalo de 120 a 179 referentes aos registros as estimativas calculadas pela MLP apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada. A Tabela 23 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absenteísmo mensal e o estimado pela MLP distribuído por classes.

Tabela 23 – Distribuição da frequência dos acertos por classes (mensal)

Acerto (mensal)	Registros	Percentual
Acima de 90%	265	80,50%
81% a 90%	47	14,30%
71% a 80,9%	6	1,80%
61% a 70,9%	0	0,00%
51% a 60,9%	11	3,30%
Total	329	100%

Observa-se que 265 registros ou 80,5% da estimativa calculada pela MLP apresenta acerto acima de 90%. Verifica-se também que 11 registros ou 3,3% apresentaram acerto entre 51% a 60%, ou seja, pior estimativa calculada pela MLP.

A Figura 49 mostra a fase de teste da MLP (Fase 1 – Etapa 1) para a previsão do absenteísmo trimestral medido em horas. O tempo de processamento foi de 17 minutos e 48 segundos.

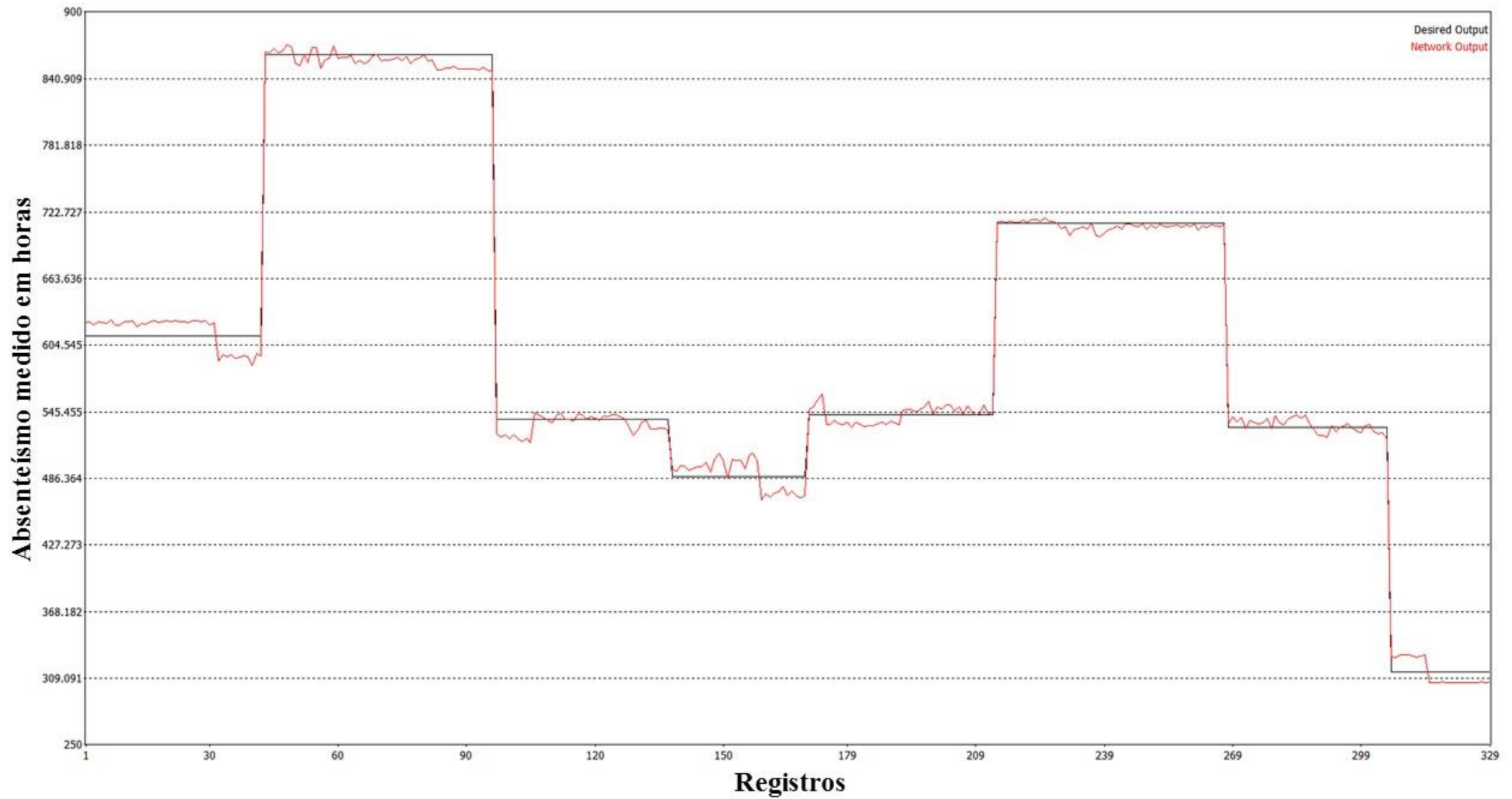


Figura 49 – Fase de teste da MLP para a previsão do absentéismo **trimestral** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0116904677 e na fase de teste igual a 0,0125162287. Percebe-se que nos intervalos de 30 a 60, 90 a 120, 150 a 179 e de 299 a 329 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada. A Tabela 24 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absenteísmo trimestral e o estimado pela MLP distribuído por classes.

Tabela 24 – Distribuição da frequência dos acertos por classes (trimestral)

Acerto (trimestral)	Registros	Percentual
Acima de 98,9%	195	59,30%
98% a 98,9%	69	21,00%
97% a 97,9%	46	14,00%
96% a 96,9%	13	4,00%
95% a 95,9%	6	1,80%
Total	329	100,00%

Observa-se que 195 registros ou 59,3% da estimativa calculada pela MLP apresenta acerto acima de 98,9%. Verifica-se também que 6 registros ou 1,8% apresentaram acerto entre 95% a 95,9%, ou seja, pior estimativa calculada.

A Figura 50 mostra a fase de teste da MLP (Fase 1 – Etapa 1) para a previsão do absenteísmo quadrimestral medido em horas. O tempo de processamento foi de 17 minutos e 48 segundos.

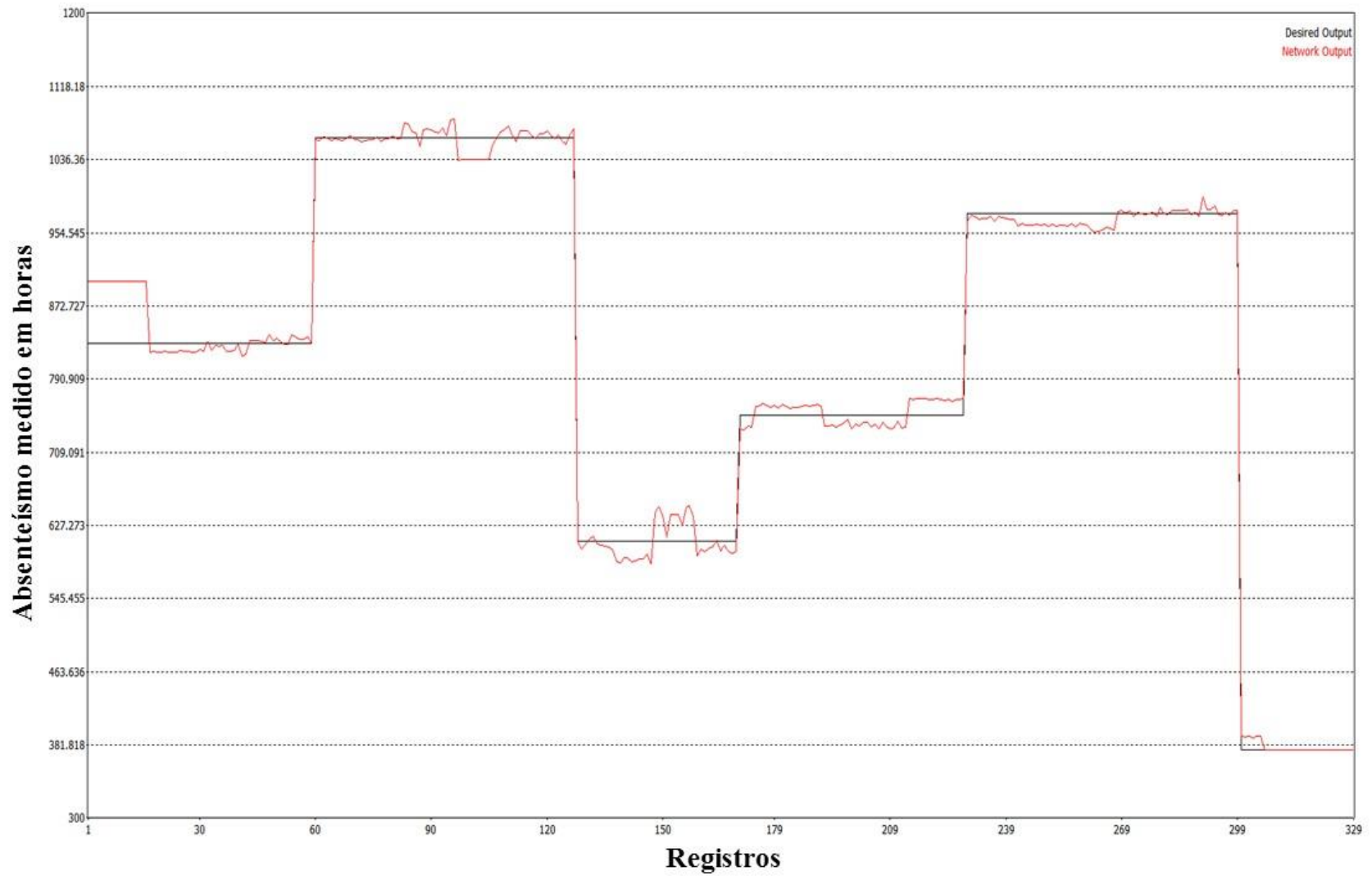


Figura 50 – Fase de teste da MLP para a previsão do absenteísmo **quadrimestral** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0116904677 e na fase de teste igual a 0,0125162287. Percebe-se que nos intervalos de 1 a 30, 90 a 120 e de 150 a 179 referentes aos registros as estimativas calculadas pela MLP apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada. A Tabela 25 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absentéismo quadrimestral e o estimado pela MLP distribuído por classes.

Tabela 25 – Distribuição da frequência dos acertos por classes (quadrimestral)

Acerto (quadrimestral)	Registros	Percentual
Acima de 97,9%	287	87,20%
96% a 97,9%	17	5,20%
94% a 95,9%	7	2,10%
92% a 93,9%	2	0,60%
90% a 91,9%	16	4,90%
Total	329	100,00%

Observa-se que 287 registros ou 87,2% da estimativa calculada pela MLP apresenta acerto acima de 97,9%. Verifica-se também que 16 registros ou 4,9% apresentaram acerto entre 90% a 91,9%, ou seja, pior estimativa calculada.

A Figura 51 mostra a curva do erro quadrático médio na fase de treinamento e teste, o eixo x possui o intervalo de 0 a 8.500 e corresponde ao número de épocas. O erro quadrático médio na fase de treinamento foi igual a 0,0116904677 e na fase de teste igual a 0,0125162287. A curva de treinamento na cor preta e a curva de teste na cor vermelha.

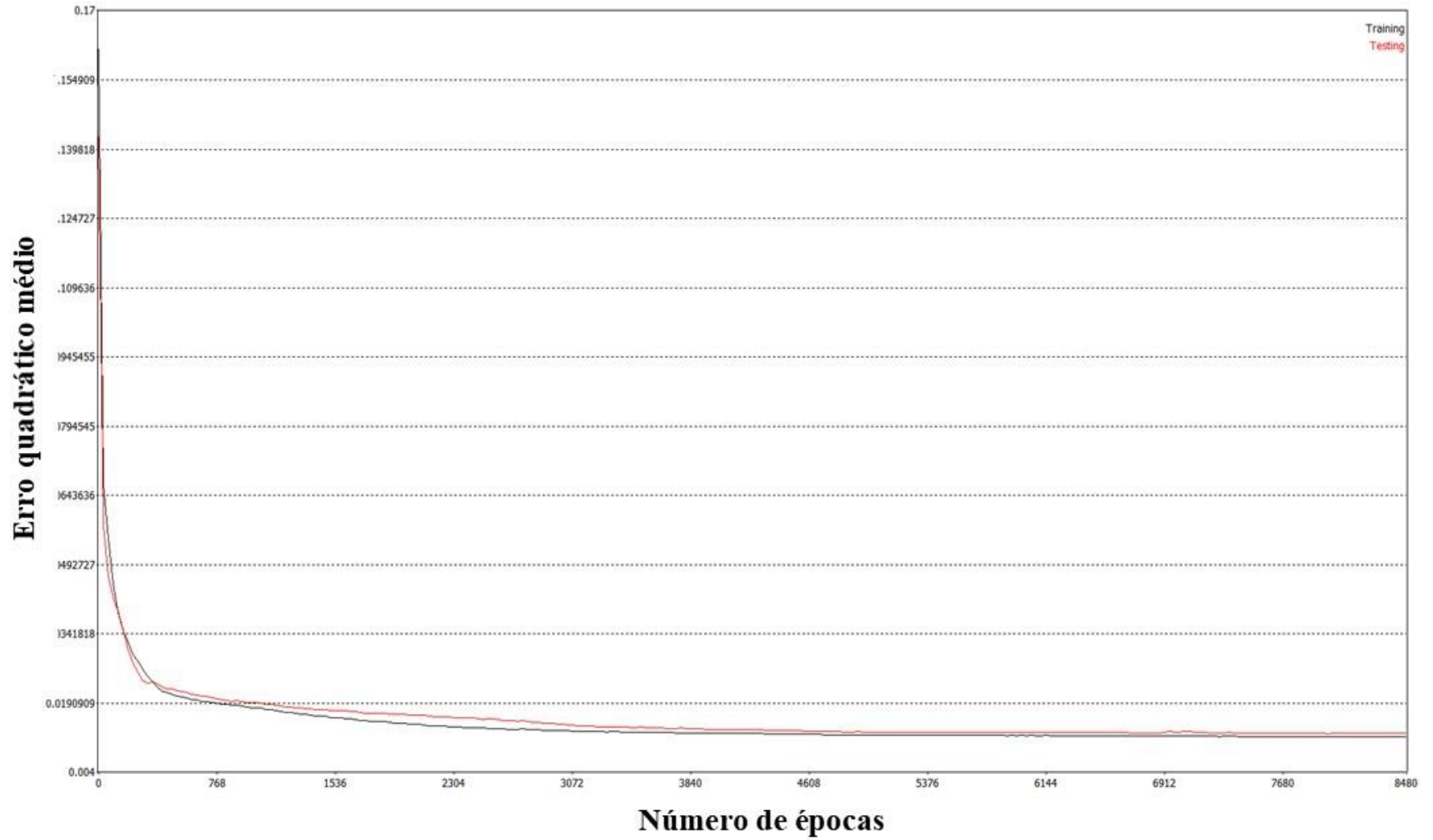


Figura 51 – Curva do erro quadrático médio na fase de **treinamento e teste**.

Observa-se que a MLP na fase de treinamento e teste apresentou o erro quadrático médio baixo, o que indica capacidade de generalizar os dados de saída desejados. Percebe-se que não ocorreu oscilação durante a fase treinamento e teste.

4.7 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos *Rough Sets* com método de redução Algoritmo genético e MLP na previsão do absentéismo (Experimento 1 - Teste).

A Figura 52 mostra a fase de teste da MLP (Fase 1, Etapa 2, Experimento 1) para a previsão do absentéismo **mensal** medido em horas. O tempo de processamento foi de 10 minutos e 40 segundos.

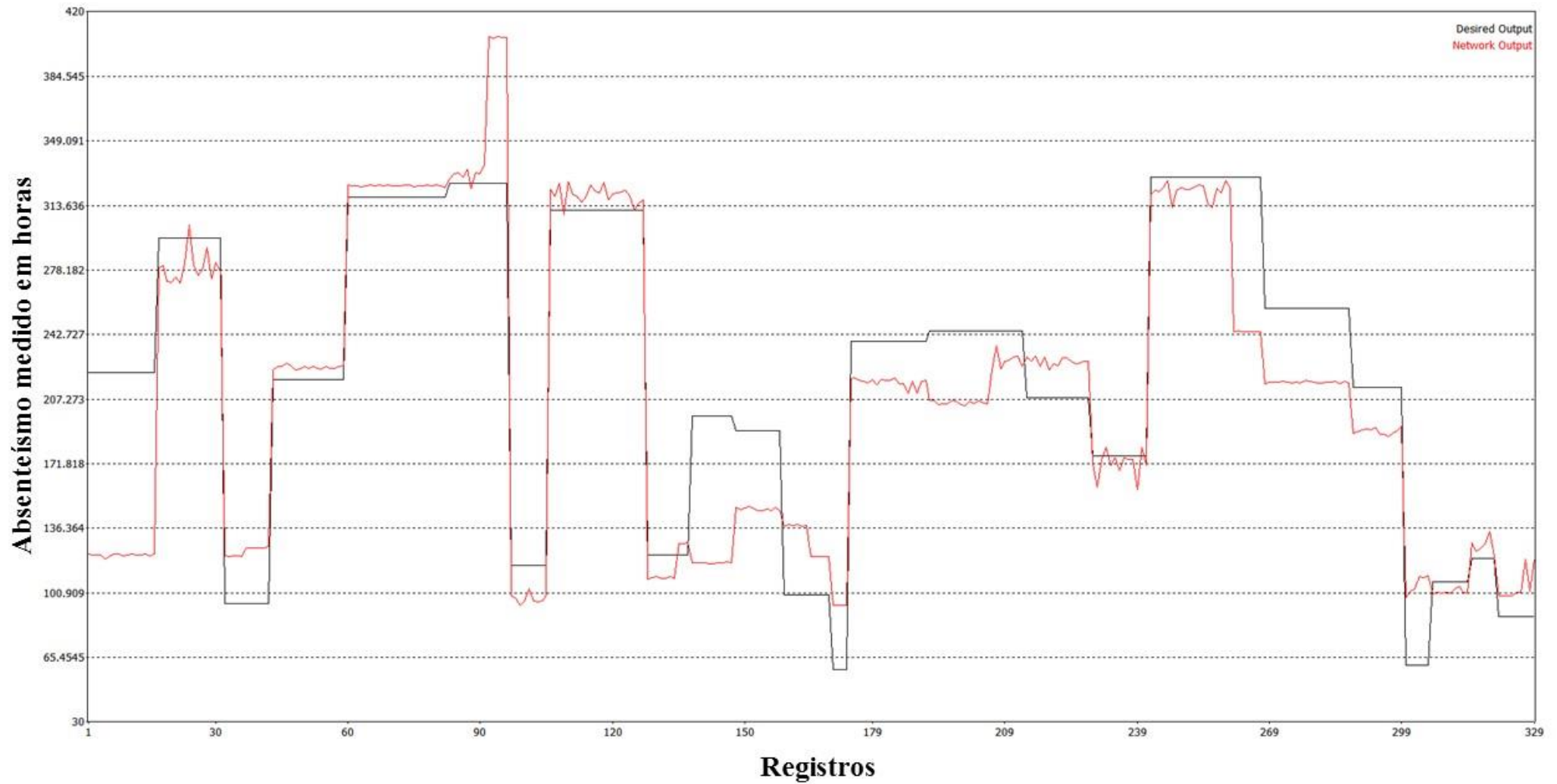


Figura 52 – Fase de teste da MLP para a previsão do absenteísmo **mensal** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0438982725 e na fase de teste foi igual a 0,0473871528. Percebe-se que nos intervalos de 1 a 30, 90 a 120 e de 120 a 179 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Tabela 26 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absentismo mensal e o estimado distribuído por classes.

Tabela 26 – Distribuição da frequência dos acertos por classes (mensal)

Acerto (mensal)	Registros	Percentual
Acima de 90%	175	53,2%
81% a 90%	73	22,2%
71% a 80%	33	10,0%
61% a 70%	14	4,3%
51% a 60%	24	7,3%
41% a 50%	4	1,2%
31% a 40%	3	0,9%
21% a 30%	2	0,6%
10% a 20%	1	0,3%
Total	329	100,0%

Observa-se que 175 registros ou 53,2% da estimativa calculada pela MLP apresenta acerto acima de 90%. Verifica-se também que apenas 1 registro ou 0,3% apresenta acerto entre 10% a 20%, ou seja, foi a pior estimativa calculada.

A Figura 53 mostra a fase de teste da MLP (Fase 1, Etapa 2, Experimento 1) para a previsão do absentismo **trimestral** medido em horas. O tempo de processamento foi de 10 minutos e 40 segundos.

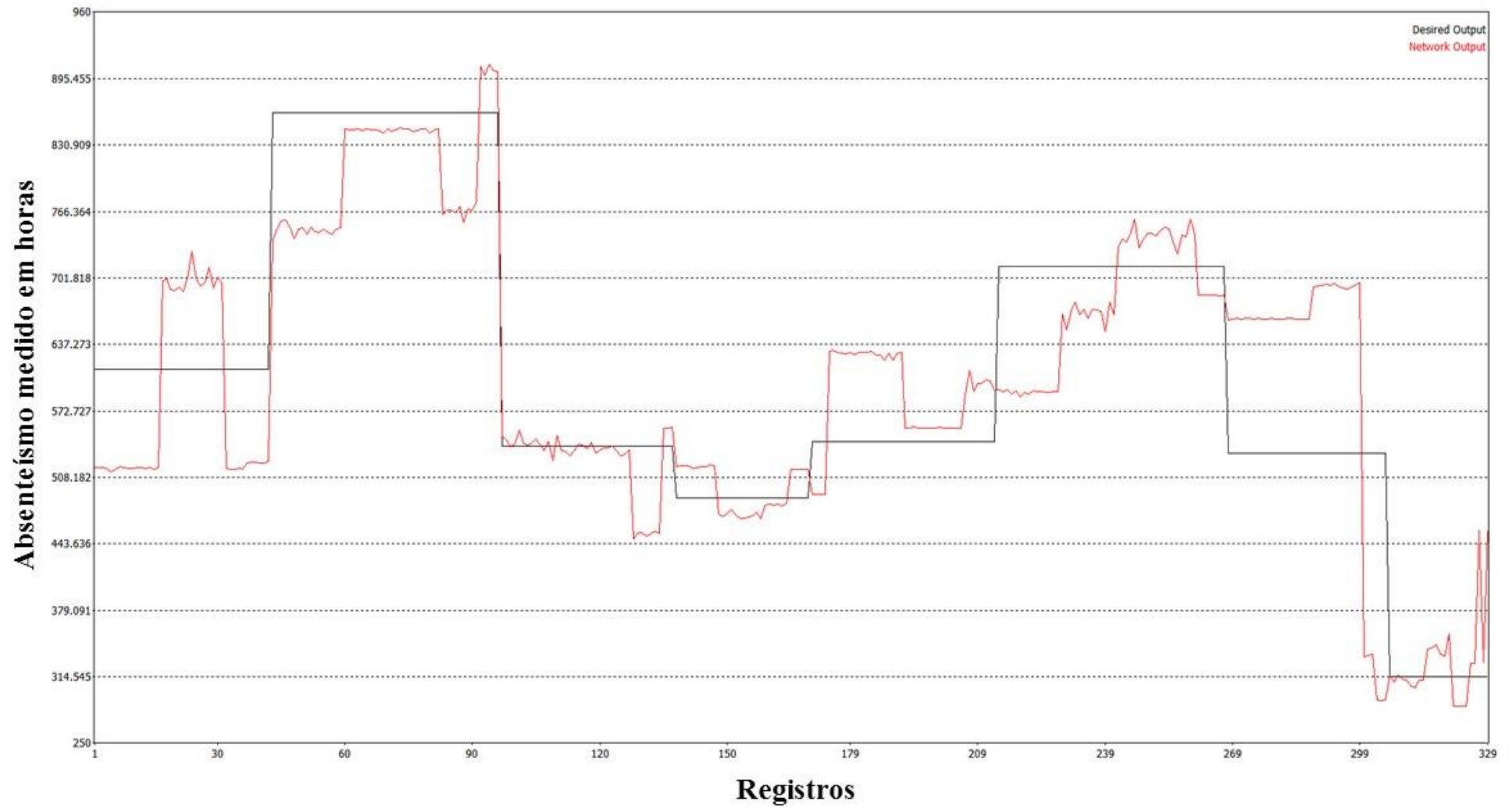


Figura 53 – Fase de teste da MLP para a previsão do absenteísmo **trimestral** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0438982725 e na fase de teste foi igual a 0,0473871528. Percebe-se que nos intervalos de 1 a 30, 30 a 90, 179 a 209 e de 269 a 299 referentes aos registros as estimativas calculadas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Tabela 27 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absenteísmo trimestral medido em horas e o estimado pela MLP distribuído por classes.

Tabela 27 – Distribuição da frequência dos acertos por classes (trimestral)

Acerto (trimestral)	Registros	(%)
Acima de 90%	178	54,1%
81% a 90%	111	33,7%
71% a 80%	26	7,9%
61% a 70%	9	2,7%
50% a 60%	5	1,5%
Total	329	100,0%

Observa-se que 178 registros ou 54,1% da estimativa calculada pela MLP apresenta acerto acima de 90%. Verifica-se também que 5 registros ou 1,5% apresentaram acerto entre 50% a 60%, ou seja, foi a pior estimativa calculada.

A Figura 54 mostra a fase de teste da MLP (Fase 1, Etapa 2, Experimento 1) para a previsão do absenteísmo **quadrimestral** medido em horas. O tempo de processamento foi de 10 minutos e 40 segundos.

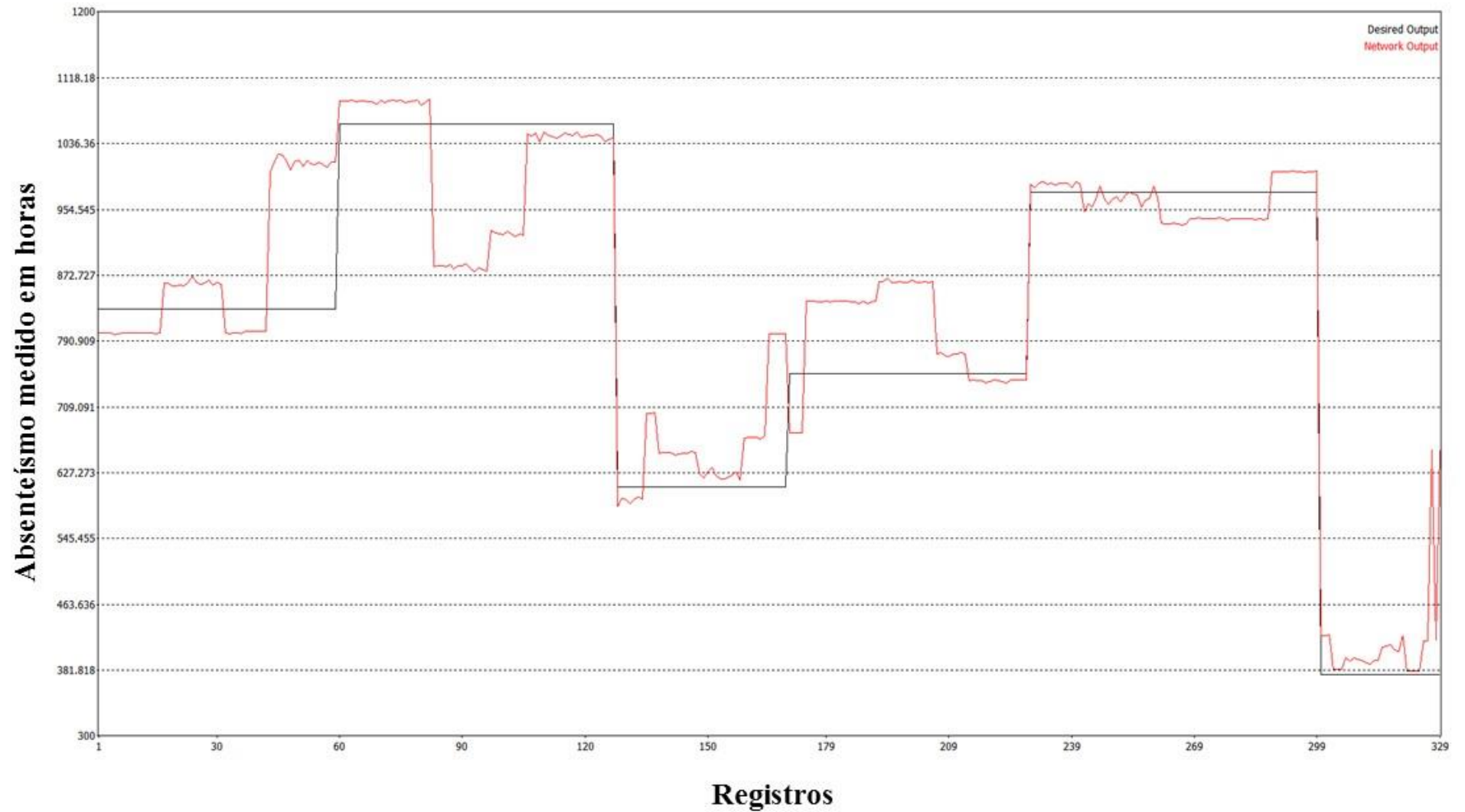


Figura 54 – Fase de teste da MLP para a previsão do absenteísmo **quadrimestral** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0438982725 e na fase de teste foi igual a 0,0473871528. Percebe-se que nos intervalos de 30 a 60, 90 a 120 e de 179 a 209 referentes aos registros as estimativas calculadas pela RNA apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Tabela 28 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absenteísmo quadrimestral medido em horas e o estimado distribuído por classes.

Tabela 28 – Distribuição da frequência dos acertos por classes (quadrimestral)

Acerto (quadrimestral)	Registros	Percentual
Acima de 90%	240	72,90%
81% a 90%	66	20,10%
71% a 80%	16	4,90%
61% a 70%	5	1,50%
50% a 60%	2	0,60%
Total	329	100,00%

Observa-se que 240 registros ou 72,9% da estimativa calculada pela MLP apresenta acerto acima de 90%. Verifica-se também que somente 2 registros ou 0,6% apresentaram acerto entre 50% a 60%, ou seja, foi a pior estimativa calculada.

A Figura 55 mostra a curva do erro quadrático médio na fase de treinamento e teste, o eixo x possui o intervalo de 0 a 8.500 e corresponde ao número de épocas da MLP. O erro quadrático médio na fase de treinamento foi igual a 0,0438982725 e na fase de teste foi igual a 0,0473871528. A curva de treinamento na cor preta e a curva de teste na cor vermelha.

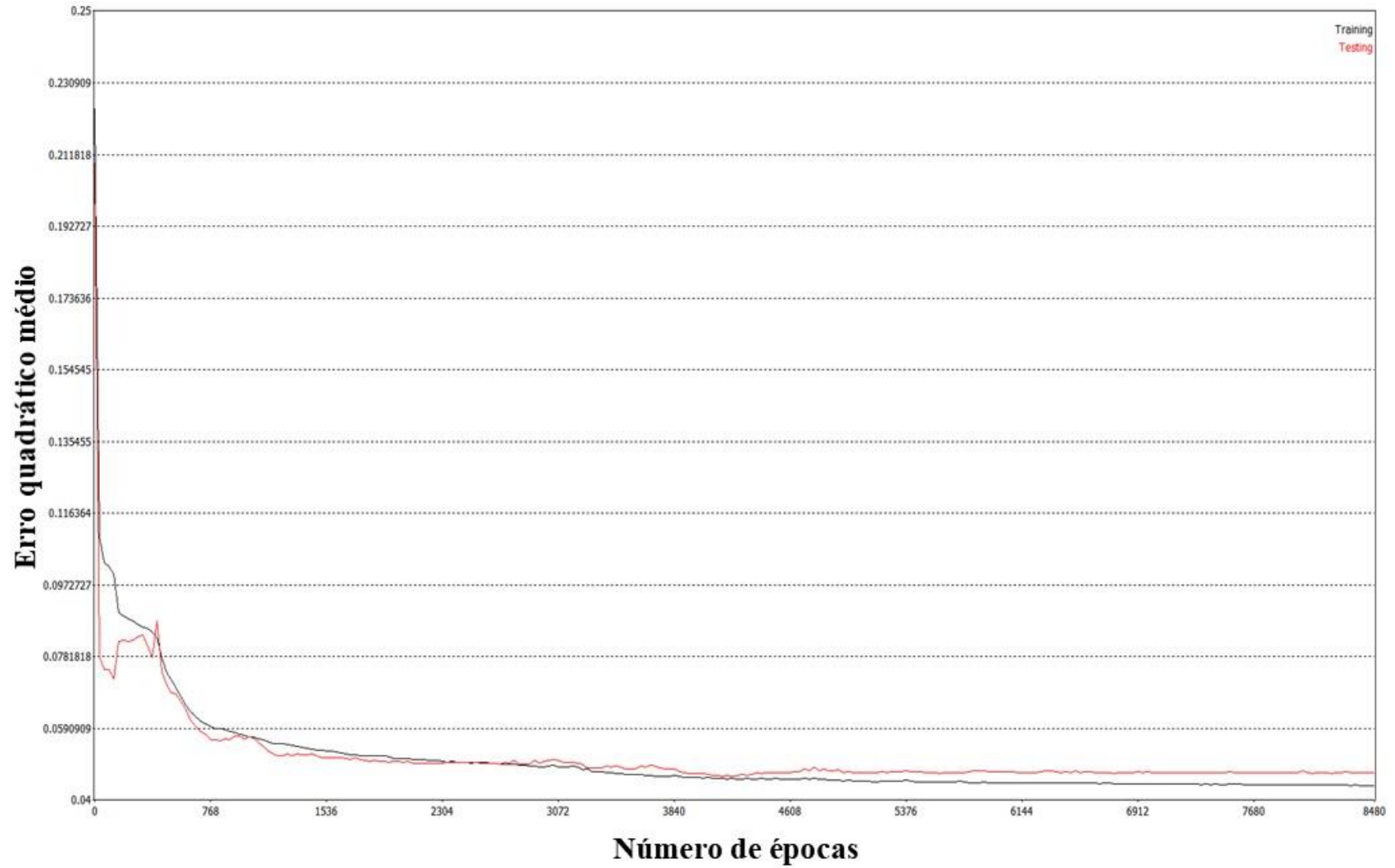


Figura 55 – Curva do erro quadrático médio na fase de **treinamento e teste**.

Observa-se que a MLP na fase de treinamento e teste apresentou o erro quadrático médio pequeno, o que indica a capacidade de generalizar os dados de saída desejados. Porém, percebe-se que ocorreu uma oscilação forte na curva do erro quadrático médio na fase de treinamento e teste no intervalo de 0 a 768 épocas e no decorrer do treinamento a oscilação diminuiu o que pode indicar ruídos característicos da base de dados utilizada.

4.8 Fase 1 - Etapa 2 (RS + RNA): Aplicação dos *Rough Sets* com método de redução Algoritmo de Johnson e MLP na previsão do absentismo (Experimento 2 - Teste).

A Figura 56 mostra a fase de teste da MLP Fase 1, Etapa 2, Experimento 2 para a previsão do absentismo **mensal** medido em horas. A curva na cor preta é a saída desejada e a curva na cor vermelha a saída da MLP. O tempo de processamento foi de 9 minutos e 13 segundos.

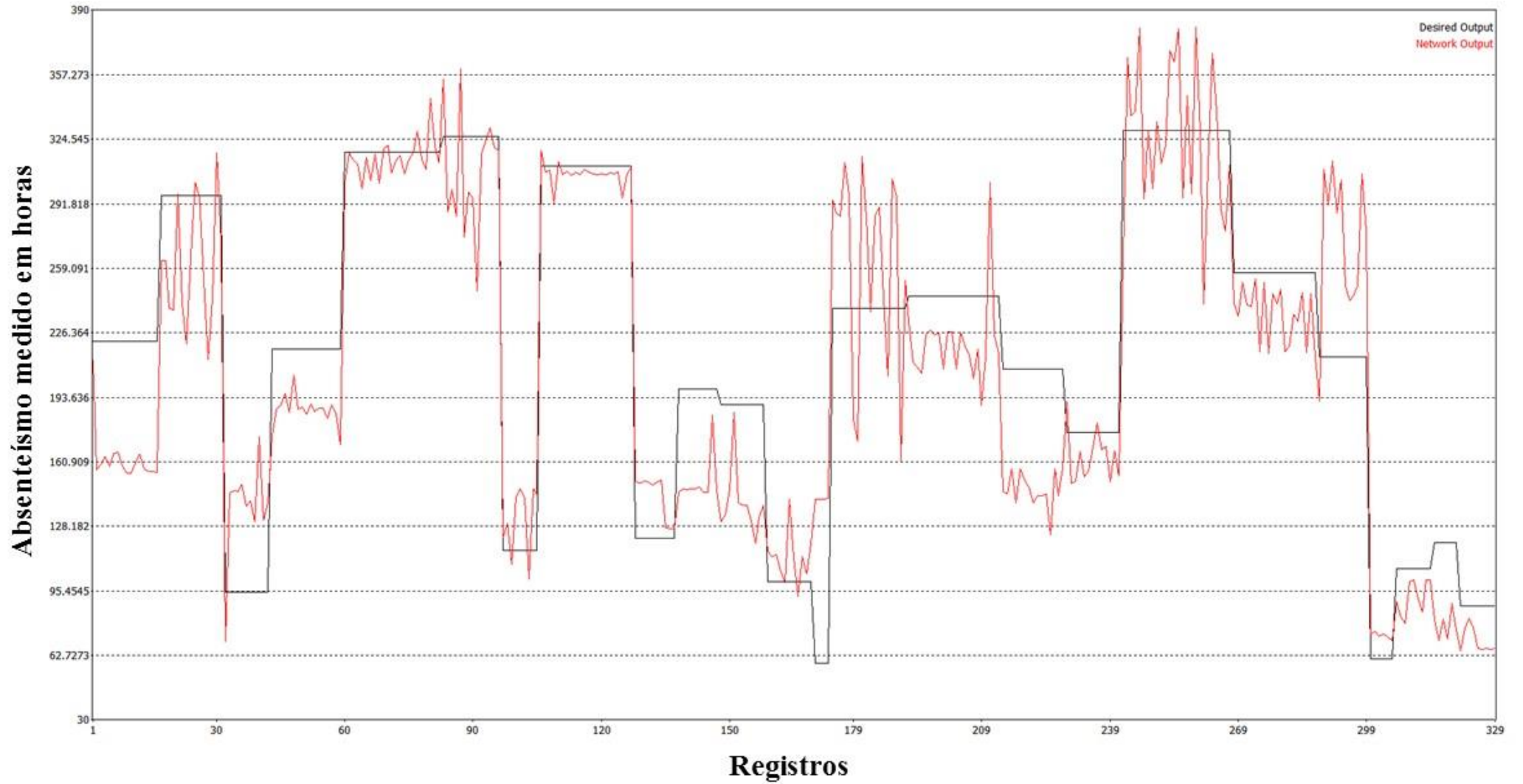


Figura 56 – Fase de teste da MLP para a previsão do absenteísmo **mensal** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0535242598 e na fase de teste foi igual a 0,0553627328. Percebe-se que nos intervalos de 1 a 30, 179 a 239 e de 269 a 299 referentes aos registros as estimativas calculadas pela MLP apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Tabela 29 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absenteísmo mensal medido em horas e o estimado pela MLP distribuído por classes.

Tabela 29 – Distribuição da frequência dos acertos por classes (mensal)

Acerto (mensal)	Registros	Percentual
Acima de 90%	128	38,90%
81% a 90%	83	25,20%
71% a 80%	80	24,30%
61% a 70%	19	5,80%
51% a 60%	10	3,00%
41% a 50%	4	1,20%
31% a 40%	0	0,00%
21% a 30%	1	0,30%
10% a 20%	4	1,20%
Total	329	100,00%

Observa-se que 128 registros ou 38,9% da estimativa calculada pela MLP apresenta acerto acima de 90%. Verifica-se também que somente 4 registros ou 1,2% apresentaram acerto entre 10% a 20%, ou seja, foi a pior estimativa calculada.

A Figura 57 mostra a fase de teste da MLP Fase 1, Etapa 2, Experimento 2 para a previsão do absenteísmo **trimestral** medido em horas. O tempo de processamento foi de 9 minutos e 13 segundos.

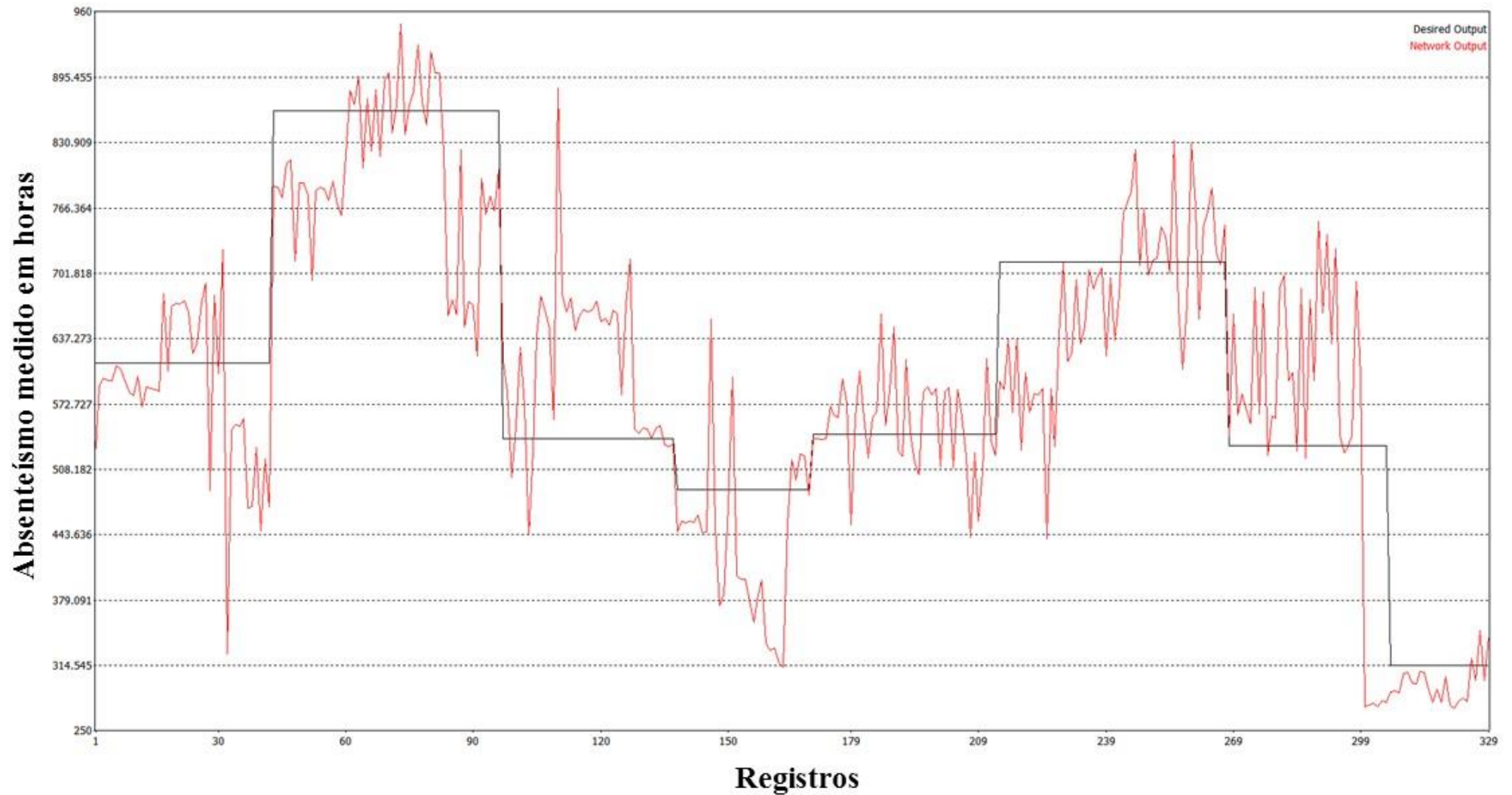


Figura 57 – Fase de teste da MLP para a previsão do absenteísmo **trimestral** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0535242598 e na fase de teste foi igual a 0,0553627328. Percebe-se que nos intervalos de 30 a 60, 90 a 120, 150 a 179, 209 a 239 e de 269 a 299 referentes aos registros as estimativas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Tabela 30 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absenteísmo trimestral medido em horas e o estimado pela MLP distribuído por classes.

Tabela 30 – Distribuição da frequência dos acertos por classes (trimestral)

Acerto (trimestral)	Registros	Percentual
Acima de 90%	201	61,10%
81% a 90%	62	18,80%
71% a 80%	46	14,00%
61% a 70%	11	3,30%
51% a 60%	8	2,40%
41% a 50%	0	0,00%
30% a 40%	1	0,30%
Total	329	100,00%

Observa-se que 201 registros ou 61,1% da estimativa calculada pela MLP apresenta acerto acima de 90%. Verifica-se também que somente 1 registro ou 0,3% apresenta acerto entre 30% a 40%, ou seja, foi a pior estimativa calculada. Verifica-se ainda que o percentual de acerto acima de 90% foi maior que o percentual acerto obtido no Experimento 1.

A Figura 58 mostra a fase de teste da MLP Fase 1, Etapa 2, Experimento 2 para a previsão do absentismo **quadrimestral** medido em horas. O tempo de processamento foi de 9 minutos e 13 segundos.

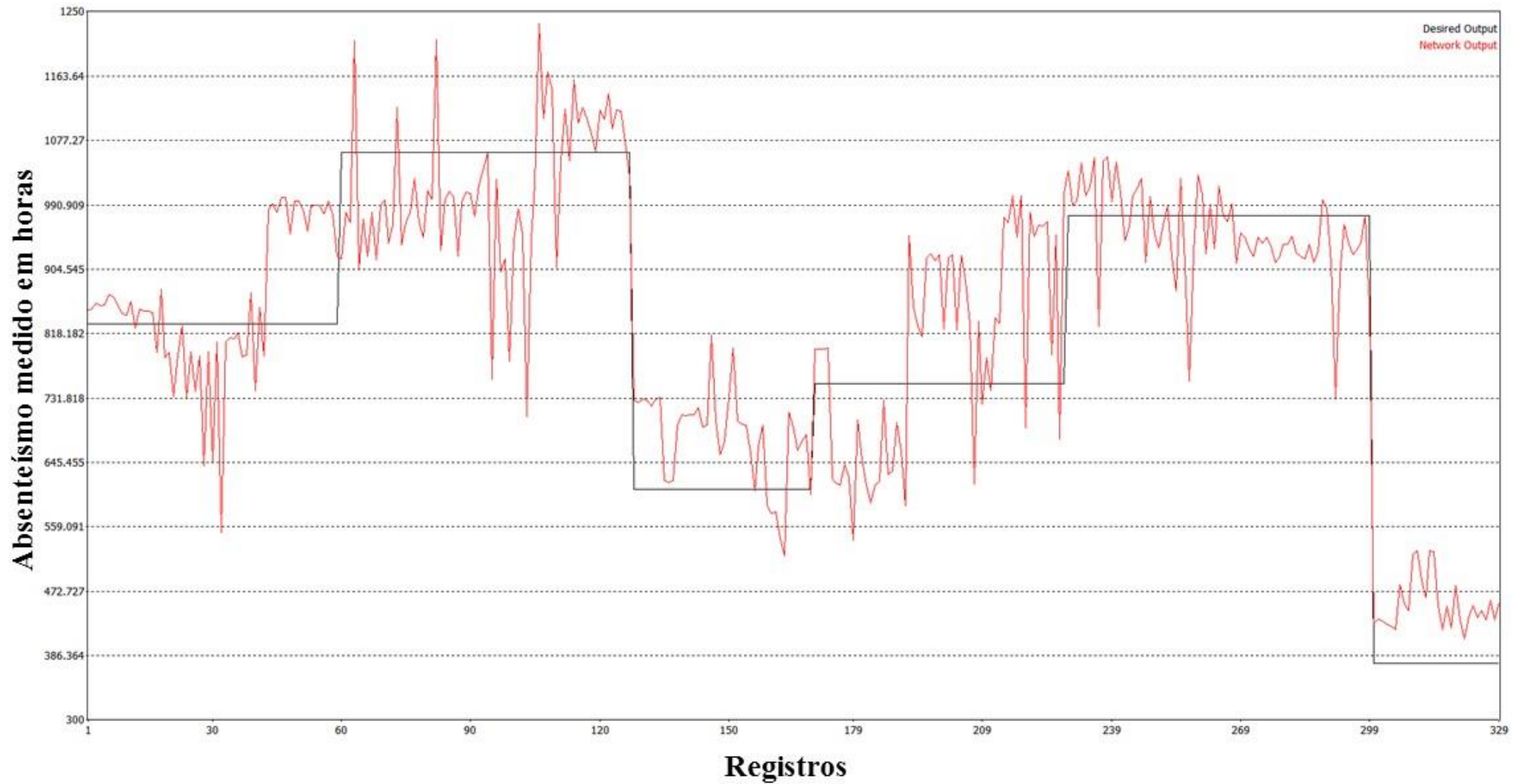


Figura 58 – Fase de teste da MLP para a previsão do absenteísmo **quadrimestral** medido em horas.

Observa-se que a MLP na fase de teste conseguiu representar os dados de saída desejados. Tendo em vista que, o erro quadrático médio na fase de treinamento foi igual a 0,0535242598 e na fase de teste foi igual a 0,0553627328. Percebe-se que nos intervalos de 30 a 60, 90 a 120 e de 179 a 239 referentes aos registros as estimativas apresentaram menor aderência, provavelmente devido aos ruídos característicos da base de dados utilizada.

A Tabela 31 apresenta a quantidade de registros e a porcentagem dos acertos experimentais entre o absentéismo trimestral e o estimado pela MLP distribuído por classes.

Tabela 31 – Distribuição da frequência dos acertos por classes (quadrimestral)

Acerto (quadrimestral)	Registros	Percentual
Acima de 90%	181	55,00%
81% a 90%	102	31,00%
71% a 80%	34	10,30%
60% a 70%	12	3,60%
Total	329	100,00%

Observa-se que 181 registros ou 55,0% da estimativa calculada pela MLP apresenta acerto na previsão do absentéismo acima de 90%. Verifica-se também que 12 registros ou 3,6% apresentaram acerto entre 60% a 70%, ou seja, foi a pior estimativa calculada.

A Figura 59 mostra a curva do erro quadrático médio na fase de treinamento e teste, o eixo x possui o intervalo de 0 a 8.500 e corresponde ao número de épocas da MLP. O erro quadrático médio na fase de treinamento foi igual a 0,0535242598 e na fase de teste foi igual a 0,0553627328. A curva de treinamento na cor preta e a curva de teste na cor vermelha.

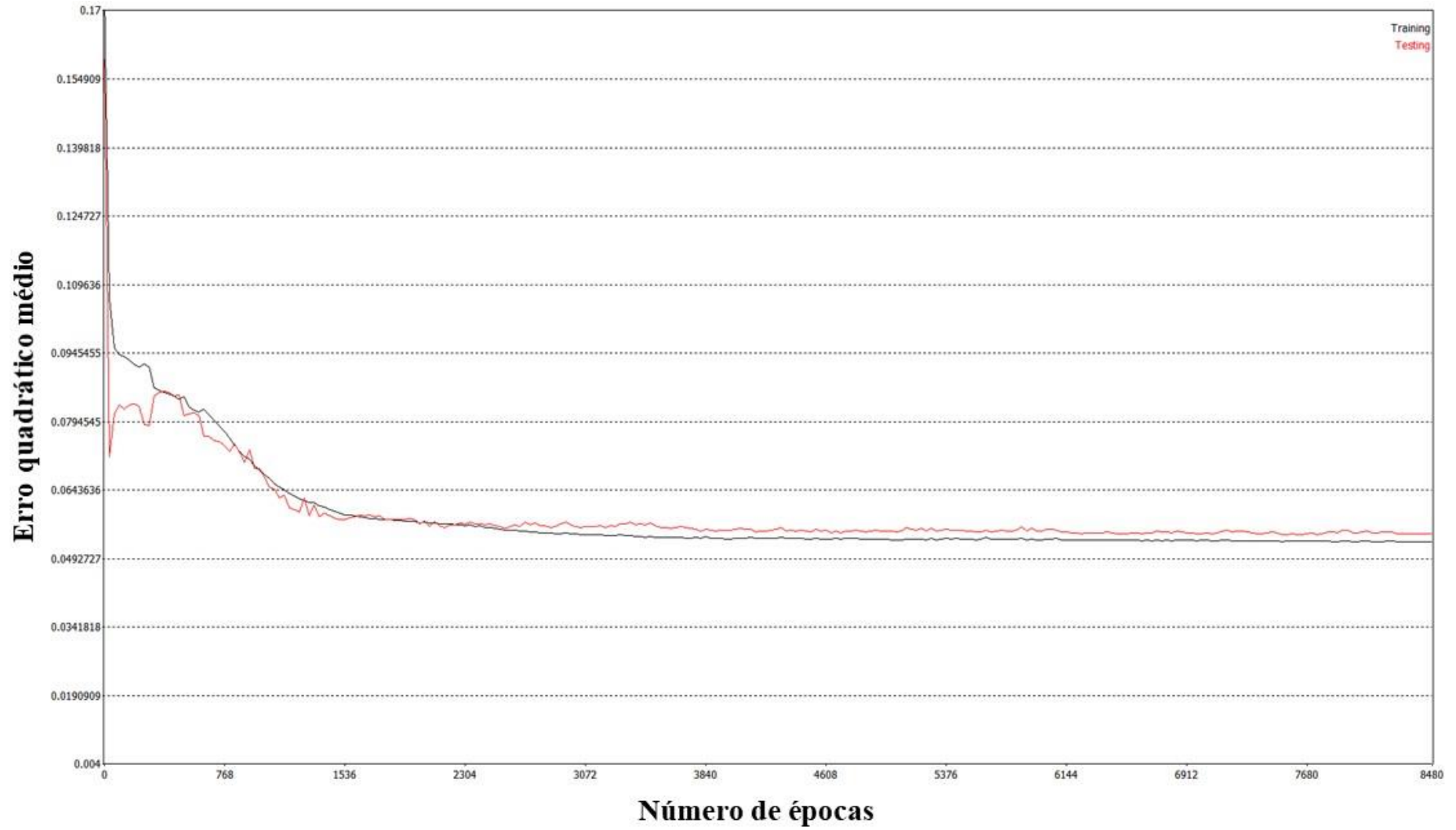


Figura 59 – Curva do erro quadrático médio na fase de **treinamento** e teste.

Observa-se que a MLP, na fase de treinamento e teste, apresentou o erro quadrático médio pequeno, o que indica a capacidade de generalizar os dados de saída desejados. Além disso, percebeu-se uma oscilação forte no intervalo de 0 a 1536 épocas.

Na seção seguinte, apresenta-se uma síntese dos resultados dos experimentos computacionais da Fase 1 (Teste).

4.9 Síntese dos resultados dos experimentos computacionais da Fase 1 – Etapas 1 e 2 (Teste)

No Quadro 6, apresenta-se uma síntese dos resultados dos experimentos computacionais da Fase 1, Etapas 1 e 2 (Teste), considerando o número de atributos, o número de entradas e de saídas da MLP, o tempo de processamento e o período.

Quadro 6 – Características dos experimentos computacionais da Fase 1 – Etapas 1 e 2 (Teste)

Experimentos Fase 1 Previsão do Absenteísmo	Número de atributos	Número de entradas	Número de saídas	Tempo de processamento	Período
Etapa 1 – RNA	50	47	3	17 minutos e 48 segundos	Mensal, trimestral, quadrimestral
Etapa 2 (RS + RNA com método de redução Algoritmo Genético) Experimento 1	20	17	3	10 minutos e 40 segundos	Mensal, trimestral, quadrimestral
Etapa 2 (RS + RNA com método de redução Algoritmo de Johnson) Experimento 2	19	16	3	9 minutos e 13 segundos	Mensal, trimestral, quadrimestral

Com as bases de dados reduzidas, observou-se que, nos experimentos computacionais, realizados na Etapa 2, o tempo de processamento foi menor, indicando menor custo computacional envolvido. De 30 atributos (Etapa 2 – Experimento 1) e 31 atributos (Etapa 2 – Experimento 2) dos 50 atributos foram desnecessários para se realizar a previsão. Fator este considerado importante, uma vez que o aumento vertiginoso das bases é uma realidade, em função da quantidade de dados produzida. Além da redução do custo computacional, considerou-se, como segundo fator importante, a não necessidade de coleta de atributos, já que

demandaria mais tempo, algo difícil de se obter junto aos empregados, e a possibilidade de armazenamento das informações, com valores faltantes e ou inconsistentes, em empresas que apresentam problemas de armazenamento.

No Quadro 7, apresenta-se uma síntese dos resultados dos experimentos computacionais da Fase 1, Etapas 1 e 2 (Teste). Nele, pode-se observar o acerto na previsão do absenteísmo por período (Etapa 1 e Etapa 2, com Experimentos 1 e 2).

Quadro 7 – Síntese dos resultados dos experimentos computacionais da Fase 1 –
Etapas 1 e 2 (Teste)

Previsão do Absenteísmo	Etapa 1 – RNA	Etapa 2 (RS + RNA com método de redução Algoritmo Genético) Experimento 1	Etapa 2 (RS + RNA com método de redução Algoritmo de Johnson) Experimento 2
Previsão acima de 90% de acerto (mensal)	80,50%	53,20%	38,90%
Previsão acima de 80% de acerto (mensal)	94,80%	75,40%	64,10%
Previsão acima de 90% de acerto (trimestral)	100,00%	54,10%	61,10%
Previsão acima de 80% de acerto (trimestral)	100,00%	87,80%	79,90%
Previsão acima de 90% de acerto (quadrimestral)	100,00%	72,90%	55,00%
Previsão acima de 80% de acerto (quadrimestral)	100,00%	93,00%	86,00%

Na fase de teste da MLP (Etapa 1), observou-se que os experimentos apresentaram melhores resultados com a base de dados sem redução. No entanto, os experimentos com as bases reduzidas também apresentaram resultados interessantes, tendo em vista a complexidade da previsão.

Percebeu-se ainda que a Etapa 1 possui melhores resultados na previsão do absenteísmo mensal, trimestral e quadrimestral. Já na Etapa 2, o Experimento 1 mostrou melhores resultados que o Experimento 2, na previsão do absenteísmo mensal e quadrimestral. No entanto, o

Experimento 2, com acerto acima de 90%, no período trimestral, foi superior ao Experimento 1.

Ainda com base no Quadro 7, foi possível prever o absenteísmo medido em horas, nas Etapas 1 e 2, com resultados distintos, o que pode auxiliar na decisão de qual modelo adotar, dependendo dos recursos disponíveis e das condições de coleta de dados.

A continuação, os experimentos computacionais para identificação de tendências absenteístas são demonstrados.

4.10 Fase 2: Identificação de tendências absenteístas – Etapa 3 (SOM)

Os experimentos com a rede SOM foram realizados com o objetivo de identificar tendências absenteístas. Na Figura 60, ilustra-se a realização dos experimentos da Fase 2.

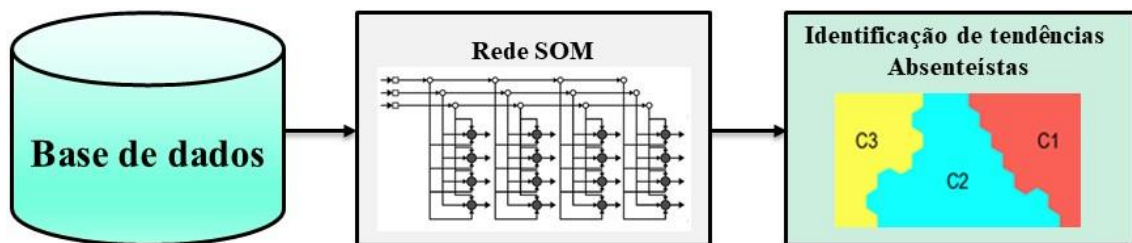


Figura 60 – Realização dos experimentos da Fase 2

Para esse experimento, foi utilizada a base de dados completa com 50 atributos. Os parâmetros usados na topologia da rede SOM foram: número de neurônios igual a 225 (15x15), vizinhança topológica hexagonal, e função de vizinhança topológica gaussiana. Os parâmetros usados na fase de treinamento foram: número de épocas (iterações) igual a 3000, taxa de aprendizado igual a 0,5, na fase inicial, e igual a 0,05, na fase de convergência (KASKI e KOHONEN, 1997). Foram gerados três agrupamentos com erro de quantização de 0,434318, e o tempo de processamento foi de 19 segundos. Na Figura 61, pode-se observar os três agrupamentos gerados pela rede SOM.

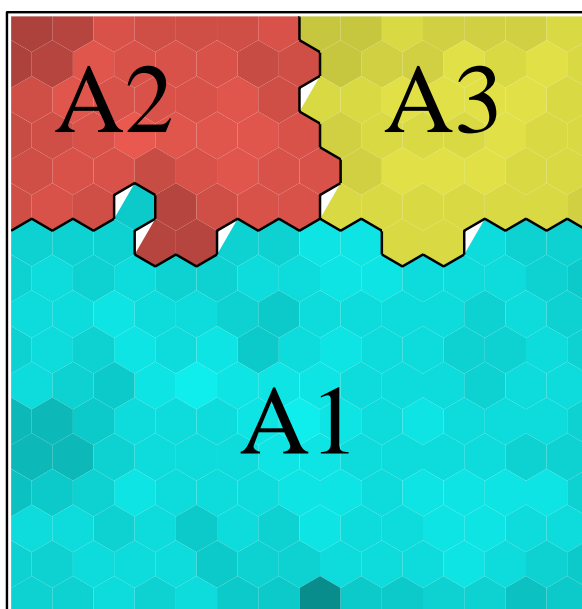


Figura 61 – Os três agrupamentos gerados pela rede SOM

Na Tabela 32, mostra-se a representatividade dos três agrupamentos gerados, considerando o número de registros da base de dados.

Tabela 32 – Representatividade dos três agrupamentos gerados pela rede SOM

Agrupamento	Registros	Percentual
A1	1524	63,42%
A2	400	16,65%
A3	479	19,93%
Total	2403	100%

Já no Quadro 8, destacam-se as características dos atributos de cada agrupamento gerado pela rede SOM, considerando o período de 2008 a 2017. Para realização da Etapa 3, 50 atributos foram considerados, ou seja, a base de dados completa.

Quadro 8 – Características dos atributos de cada agrupamento gerado pela rede SOM

Características	A1	A2	A3
Idade (média)	38,69 anos	55,38 anos	36,14 anos
Tempo de serviço (média)	14,16 anos	28,69 anos	4,09 anos
Número de filhos (média)	1,10	2,00	0,27
Motivo da ausência (acompanhante)	276 registros – 2042 horas – média 7,39 horas	51 registros – 291 horas – média 5,70 horas	3 registros – 20 horas – média 6,66 horas
CID Capítulo 13 – Doenças do sistema osteomuscular e do tecido conjuntivo (M00-M99)	145 registros – 2304 horas – média 7,39 horas	63 registros – 1297 horas – média 20,58 horas	19 registros – 169 horas – média 8,89 horas
CID Capítulo 19 – Lesões, envenenamento e algumas outras consequências de causas externas (S00-T98)	91 registros – 2483 horas – média 27,28 horas	25 registros – 845 horas – média 33,8 horas	5 registros – 57 horas – média 11,4 horas

Os agrupamentos A1 e A3 apresentam os empregados mais jovens, com média de idade de 38,69 e 36,14 anos, respectivamente. Contudo, existe uma diferença, de mais de 10 anos, em média, no tempo de serviço.

O agrupamento A2 apresenta os empregados mais idosos, com média de idade de 55,38 anos. Conforme demonstrado na Tabela 33, as ausências relativas ao CID Capítulo 13, ou seja, doenças do sistema osteomuscular e do tecido conjuntivo (M00-M99), apresentam a maior média em horas de absenteísmo, o que pode indicar problemas relacionados ao trabalho em si, ao tempo de serviço e à idade.

As ausências pelo CID Capítulo 19, isto é, lesões, envenenamento e algumas outras consequências de causas externas (S00-T98), são mais significativas no agrupamento A1 e A2, e apresentam as maiores médias de absenteísmo, 27,28 horas e 33,80 horas, respectivamente. Isso pode indicar agrupamentos com maior incidência de pequenos acidentes com afastamentos maiores que 24 horas ou 3 dias, em média.

Quanto à média do tempo de serviço, observou-se também que os agrupamentos A1 e A2 são bem distintos, sendo que o agrupamento A2 apresenta empregados com o dobro de média do tempo de serviço em comparação ao A1.

Quanto ao número de filhos, o agrupamento A2 apresenta empregados com maior média (2 filhos), contudo, é o agrupamento A1 que se destaca com maior quantidade de horas de

afastamento pelo motivo “acompanhante”, sendo 2042 horas, e maior média (7,39 horas) por registro.

A seguir, apresenta-se uma análise da tendência do absenteísmo, com dois períodos distintos: (1) anual, durante um período de dez anos; e (2) bianual, entre o período de 2008 a 2017, ou seja, cinco biênios. Os resultados dessas análises foram extraídos dos agrupamentos gerados pela rede SOM.

a) Análise da tendência do absenteísmo anual

Nas Tabelas 33 a 35, constam informações da tendência do absenteísmo anual. Cada agrupamento foi verificado quanto à redução da frequência ou ao aumento da frequência das licenças médicas. No ano de 2008, por exemplo, as frequências das licenças médicas diminuíram no agrupamento A1. Isso significa que os empregados desse agrupamento deixaram de apresentar licenças médicas com determinada frequência (redução) no respectivo ano.

Desta forma, constatou-se uma redução da frequência. Vale ressaltar que essa redução independe do absenteísmo medido em horas. Um exemplo disso é que, no ano de 2013, mesmo com 1847 horas, ocorreu uma redução da frequência no agrupamento, em um grupo específico de empregados. Na Tabela 33, pode-se observar a tendência do absenteísmo anual no agrupamento A1.

Tabela 33 – Tendência do absenteísmo anual no agrupamento A1

Agrupamento A1 – Análise anual		
Ano	Absenteísmo medido em horas	Tendência
2008	82	Redução da frequência
2009	45	Redução da frequência
2010	967	Redução da frequência
2011	1627	Redução da frequência
2012	1421	Redução da frequência
2013	1847	Redução da frequência
2014	1665	Redução da frequência
2015	2008	Aumento da frequência
2016	1547	Redução da frequência
2017	1675	Aumento da frequência
Total	12884	

O agrupamento A1 mostrou possuir maior representatividade, com 63,42% dos registros, apresentando um absenteísmo, medido em horas, maior do que os demais

agrupamentos. No agrupamento A1, verificou-se uma tendência ao aumento da frequência do absenteísmo nos anos de 2015 e 2017.

Esse aumento pode indicar problema organizacional, políticas internas, falta de condições adequadas de trabalho, ambiente inseguro, remuneração inadequada, e gestão ultrapassada. Na Tabela 34, pode-se observar a tendência do absenteísmo anual no agrupamento A2.

Tabela 34 – Tendência do absenteísmo anual no agrupamento A2

Agrupamento A2 – Análise anual		
Ano	Absenteísmo medido em horas	Tendência
2008	321	Redução da frequência
2009	320	Redução da frequência
2010	886	Aumento da frequência
2011	1128	Redução da frequência
2012	819	Redução da frequência
2013	932	Aumento da frequência
2014	1147	Redução da frequência
2015	382	Redução da frequência
2016	965	Aumento da frequência
2017	428	Redução da frequência
Total	7328	

O agrupamento A2 apresentou a menor representatividade, com 16,65% dos registros. Observou-se uma tendência ao aumento da frequência do absenteísmo nos anos de 2010, 2013 e 2016. Tal aumento pode indicar problema organizacional, políticas internas, falta de condições adequadas de trabalho, ambiente inseguro, remuneração inadequada, e gestão ultrapassada. Na Tabela 35, pode-se observar a tendência do absenteísmo anual no agrupamento A3.

Tabela 35 – Tendência do absenteísmo anual no agrupamento A3

Agrupamento A3 - Análise anual		
Ano	Absenteísmo medido em horas	Tendência
2008	1234	Aumento da frequência
2009	880	Redução da frequência
2010	484	Redução da frequência
Total	2598	

O A3 foi o segundo agrupamento a possuir maior representatividade, com 19,93% dos registros, apresentando uma tendência ao aumento da frequência do absenteísmo no ano de 2008 (totalizando 1234 horas), não possuindo registros entre 2011 e 2017.

Esse tipo de aumento pode indicar problema organizacional, políticas internas, falta de condições adequadas de trabalho, ambiente inseguro, remuneração inadequada, e gestão ultrapassada.

Observou-se que os três agrupamentos apresentaram um aumento da frequência do absenteísmo em anos distintos. O agrupamento A1 apresentou aumento da frequência em 2015 e 2017, o agrupamento A2 apresentou aumento da frequência em 2013 e 2016, e o agrupamento A3, apenas em 2008. Infelizmente, os motivos que levaram à redução e ao aumento da frequência do absenteísmo não puderam ser investigados, devido à falta de dados.

b) Análise da tendência do absenteísmo bianual

Nas Tabelas 36 a 38, constam informações da tendência do absenteísmo bianual. Cada agrupamento foi verificado quanto à redução da frequência ou ao aumento da frequência das licenças médicas. No biênio 2008-2009, por exemplo, as frequências das licenças médicas reduziram no agrupamento A1. Isso significa que o grupo de empregados desse agrupamento deixaram de apresentar licenças médicas com determinada frequência (redução) no respectivo biênio.

Desta forma, conforme demonstrado na Tabela 36, constatou-se uma redução da frequência. Vale ressaltar que essa redução independe do absenteísmo medido em horas. Como exemplo disso, pode-se citar o biênio 2014-2015 que, mesmo com 3673 horas, apresentou uma redução da frequência em um grupo específico de empregados. Na Tabela 36, é possível observar a tendência do absenteísmo bianual no agrupamento A1.

Tabela 36 – Tendência do absenteísmo bianual no agrupamento A1

Agrupamento A1 – Análise bianual		
Ano	Absenteísmo medido em horas	Tendência
2008 - 2009	127	Redução da frequência
2010 - 2011	2594	Redução da frequência
2012 - 2013	3268	Redução da frequência
2014 - 2015	3673	Redução da frequência
2016 - 2017	3222	Aumento da frequência
Total	12884	

O agrupamento A1 apresentou uma tendência ao aumento da frequência do absenteísmo no período bianual 2016-2017. Isso indica que 2017 foi o ano em que ocorreu um aumento da frequência acentuado, já que o ano de 2016, de maneira isolada, apresentou uma redução da frequência, conforme demonstrado na análise anual (Tabela 33). Já no período bianual 2014-2015, ocorreu o inverso, ou seja, em 2014, a redução da frequência foi acentuada, já que o ano de 2015, de maneira isolada, apresentou aumento da frequência, conforme demonstrado na análise anual do agrupamento A1 (Tabela 33).

Diante disso, o agrupamento A1 mostrou possuir maior representatividade, o que pode indicar um problema organizacional, como descrito anteriormente, causando, assim, o aumento da frequência de absenteísmo. Desta forma, o ano de 2017 merece uma investigação mais cuidadosa quanto aos prováveis motivos do absenteísmo. Na Tabela 37, pode-se observar a tendência do absenteísmo bianual no agrupamento A2.

Tabela 37 – Tendência do absenteísmo bianual no agrupamento A2

Agrupamento A2 – Análise bianual		
Ano	Absenteísmo medido em horas	Tendência
2008 - 2009	641	Redução da frequência
2010 - 2011	2014	Redução da frequência
2012 - 2013	1751	Aumento da frequência
2014 - 2015	1529	Redução da frequência
2016 - 2017	1393	Redução da frequência
Total	7328	

Quanto ao agrupamento A2, percebeu-se uma tendência ao aumento da frequência do absenteísmo no período bianual 2012-2013. Isso indica que 2013 foi o ano em que ocorreu um aumento da frequência acentuado, já que o ano de 2012, de maneira isolada, apresentou uma redução da frequência, conforme demonstrado na análise anual (Tabela 34). Observou-se também que, nos períodos bianuais 2010-2011 e 2016-2017, ocorreu o inverso, ou seja, nos anos de 2011 e 2017, a redução da frequência foi acentuada, uma vez que os anos de 2010 e 2016, isoladamente, apresentam aumento da frequência, como demonstrado na análise anual do agrupamento A2 (Tabela 34).

Desta forma, os períodos bianuais 2010-2011, 2012-2013 e 2016-2017 merecem uma investigação mais cuidadosa quanto às prováveis causas do absenteísmo. O agrupamento A2 pode indicar algum problema organizacional, como descrito anteriormente, o que pode causar o aumento da frequência de absenteísmo nos períodos bianuais destacados. Na Tabela 38, pode-se observar a tendência do absenteísmo bianual no agrupamento A3.

Tabela 38 – Tendência do absenteísmo bianual no agrupamento A3

Agrupamento A3 – Análise bianual		
Ano	Absenteísmo medido em horas	Tendência
2008 - 2009	2114	Aumento da frequência
Total	2114	

Já o agrupamento A3 apresentou uma tendência ao aumento da frequência do absenteísmo no período bianual 2008-2009. Isso indica que 2008 foi o ano em que ocorreu um aumento da frequência acentuado, já que o ano de 2009, isoladamente, teve uma redução da frequência, conforme demonstrado na análise anual (Tabela 35). A análise do biênio seguinte (2010-2011) não foi realizada, uma vez que o agrupamento A3 não apresentava dados em 2011. Percebeu-se que o biênio 2008-2009 concentrou 64%, ou 2114 horas, de absenteísmo no agrupamento A3. Isso pode indicar algum problema organizacional, como descrito anteriormente, no período bianual destacado.

4.11 Síntese dos resultados dos experimentos computacionais da Fase 2 – Etapa 3

No Quadro 9, apresenta-se uma síntese dos resultados dos experimentos computacionais realizados na Fase 2 (Etapa 3), no período anual.

Quadro 9 – Síntese dos resultados dos experimentos computacionais no período anual

Ano	A1	Absenteísmo medido em horas	A2	Absenteísmo medido em horas	A3	Absenteísmo medido em horas
2008	Redução da frequência	82	Redução da frequência	321	Aumento da frequência	1234
2009	Redução da frequência	45	Redução da frequência	320	Redução da frequência	880
2010	Redução da frequência	967	Aumento da frequência	886	Redução da frequência	484
2011	Redução da frequência	1627	Redução da frequência	1128	N/D	N/D
2012	Redução da frequência	1421	Redução da frequência	819	N/D	N/D
2013	Redução da frequência	1847	Aumento da frequência	932	N/D	N/D
2014	Redução da frequência	1665	Redução da frequência	1147	N/D	N/D
2015	Aumento da frequência	2008	Redução da frequência	382	N/D	N/D
2016	Redução da frequência	1547	Aumento da frequência	965	N/D	N/D
2017	Aumento da frequência	1675	Redução da frequência	428	N/D	N/D
	Total	12884		7328		2598

O agrupamento A1 apresentou o total de 12884 horas de absenteísmo, sendo 9201 dos anos de 2008, 2009, 2010, 2011, 2012, 2013, 2014 e 2016, períodos em que ocorreu uma redução da frequência. Já os anos de 2015 e 2017 apresentaram 3683 horas, períodos em que ocorreu um aumento da frequência, ou seja, $3683 \text{ horas} / 12884 \text{ horas} = 0,286$ ou 28,6%. Isso representa cerca de um terço das horas de absenteísmo no agrupamento com maior representatividade.

O agrupamento A2 apresentou o total de 7328 horas de absenteísmo, sendo 4545 dos anos de 2008, 2009, 2011, 2012, 2014, 2015 e 2017, períodos em que ocorreu uma redução da frequência. Já os anos de 2010, 2013 e 2016 apresentam 2783 horas, períodos em que ocorreu um aumento da frequência, ou seja, $2783 \text{ horas} / 7328 \text{ horas} = 0,38$ ou 38%. Isso indica cerca de dois quintos das horas de absenteísmo no agrupamento.

Por fim, o agrupamento A3 apresentou o total de 2598 horas de absenteísmo, sendo 1364 dos anos de 2009 e 2010, períodos em que ocorreu uma redução da frequência. Já o ano de 2008 apresentou 1234 horas, período em que ocorreu um aumento da frequência, ou seja, $1234 \text{ horas} / 2598 \text{ horas} = 0,475$ ou 47,5%. Isso indica cerca de um meio das horas de absenteísmo no agrupamento.

No Quadro 10, apresenta-se uma síntese dos resultados dos experimentos computacionais realizados na Fase 2 (Etapa 3), no período bianual.

Quadro 10 – Síntese dos resultados dos experimentos computacionais da Fase 2 –
Etapa 3 no período bianual

Biênio	A1	Absenteísmo medido em horas	A2	Absenteísmo medido em horas	A3	Absenteísmo medido em horas
2008-2009	Redução da frequência	127	Redução da frequência	641	Aumento da frequência	2114
2010-2011	Redução da frequência	2594	Redução da frequência	2014	N/D	N/D
2012-2013	Redução da frequência	3268	Aumento da frequência	1751	N/D	N/D
2014-2015	Redução da frequência	3673	Redução da frequência	1529	N/D	N/D
2016-2017	Aumento da frequência	3222	Redução da frequência	1393	N/D	N/D
	Total	12884		7328		2114

Conforme observado no Quadro 9, o agrupamento A1 apresentou o total de 12884 horas de absenteísmo, sendo 9662 dos biênios 2008-2009, 2010-2011, 2012-2013 e 2014-2015, períodos em que ocorreu uma redução da frequência. Já o biênio 2016-2017 apresentou 3222 horas, período em que ocorreu um aumento da frequência, ou seja, $3222 \text{ horas} / 12884 \text{ horas} = 0,33$ ou 33%. Isso indica um terço das horas de absenteísmo no agrupamento com maior representatividade.

O agrupamento A2 apresentou o total de 7328 horas de absenteísmo, sendo 5577 dos biênios de 2008-2009, 2010-2011, 2014-2015 e 2016-2017, períodos em que ocorreu uma redução da frequência. Já o biênio 2012-2013 apresentou 1751 horas, período em que ocorreu um aumento da frequência, ou seja, $1751 \text{ horas} / 7328 \text{ horas} = 0,31$ ou 31%. Isso indica cerca de um terço das horas de absenteísmo no agrupamento.

Por fim, o agrupamento A3, no biênio 2008-2009, apresentou o total de 2114 horas de absenteísmo. Não foi possível comparar o aumento e a redução da frequência do absenteísmo devido à característica do agrupamento A3, que não possui outros biênios.

A síntese apresentada comprovou que a identificação da tendência absenteísta revela o quanto a empresa perde com esse fenômeno, já que a força de trabalho é o insumo primário para a produção econômica.

No próximo capítulo, são apresentadas as considerações finais deste estudo.

5. CONCLUSÃO

Neste trabalho, foi desenvolvido um estudo sobre a aplicação de técnicas de inteligência computacional na previsão do absenteísmo e identificação de tendências absenteístas. O absenteísmo é um fenômeno definido como o não comparecimento do empregado ao trabalho, de forma habitual e com frequência regular, e, conseqüentemente, como o não cumprimento das obrigações trabalhistas, conforme o programado. Sendo assim, o absenteísmo difere da falta ao trabalho, em situação pontual isolada, em função de sua frequência e habitualidade.

O absenteísmo reduz a competitividade da empresa, uma vez que aumenta os custos e diminui a capacidade produtiva e a qualidade do trabalho. Com a realidade socioeconômica atual, o aumento da concorrência tornou-se inevitável, principalmente devido ao alto índice de desemprego, impondo, assim, grandes desafios para empresas, empregados, bem como para toda a sociedade.

Nesse sentido, o absenteísmo, relacionado ao ambiente de trabalho, foi alvo de estudos em diferentes áreas do conhecimento. Além disso, o tema da previsão do absenteísmo e da identificação de tendências absenteístas vem sendo investigado, principalmente por causa da possibilidade de se tomar certas providências antes da ocorrência do absenteísmo, permitindo, assim, um tratamento preventivo.

Uma das contribuições do tema em questão é que qualquer empresa pode usar a previsão e a identificação de tendências absenteístas para planejar e apoiar decisões que gerem resultados positivos. Deste modo, tanto a organização quanto os empregados ganham em produtividade e qualidade dos produtos e serviços, assim como em qualidade de vida no trabalho.

5.1 Contribuição dos experimentos computacionais

Para desenvolvimento deste trabalho, foram realizados experimentos computacionais em duas fases: a Fase 1, denominada Previsão do Absenteísmo, foi dividida em duas etapas. Na Etapa 1, aplicou-se a rede neural artificial do tipo *Multilayer Perceptron* (MLP), e na Etapa 2, aplicou-se a Teoria dos *Rough Sets*, para redução de atributos, com uso de dois métodos: o Algoritmo Genético e o Algoritmo de Johnson. Em seguida, aplicou-se a *Multilayer Perceptron*.

Os resultados obtidos nas duas fases foram considerados bons, em função dos baixos erros experimentais observados. Apesar da pequena vantagem para os resultados da Etapa 1,

nas duas etapas a previsão foi realizada. No entanto, os tempos de processamento computacional obtidos na Etapa 2 são consideravelmente mais baixos do que o obtido na Etapa 1, sendo que o tempo de processamento tem sua relevância, dependendo da quantidade de empregados que a empresa possui. Desta forma, constatou-se que a Etapa 2 apresentou bons resultados e o melhor ganho computacional. O fato de a Etapa 2 reduzir o tempo de processamento e evitar que a empresa colete informações desnecessárias, despendendo tempo com tratamento, pesou a seu favor.

Na Fase 2 (Etapa 3), em que se utilizou a *Self-Organizing Map* (SOM), os resultados gerados também apontaram positivamente para a identificação de tendências absenteístas por meio da avaliação de agrupamentos. Os resultados dos experimentos computacionais apresentaram o aumento da tendência ao absenteísmo em determinados anos e em grupos distintos. A síntese dos resultados evidenciou o quanto a empresa perdeu sua força de trabalho em função do absenteísmo.

Com isso, considerou-se que o objetivo geral deste trabalho foi atingido, ao prever o absenteísmo e identificar tendências absenteístas com técnicas de inteligência computacional, como as redes neurais artificiais e os *Rough Sets*, uma vez que auxiliaram no entendimento de um fenômeno muitas vezes silencioso e complexo.

Entre as vantagens da previsão do absenteísmo e da identificação de tendências absenteístas, pode-se destacar a possibilidade de se tomar determinadas medidas antes da ocorrência do absenteísmo, ou mesmo durante sua fase inicial. Isto é, certas ações podem ser implementadas para que o absenteísmo seja evitado ou tratado, de modo que suas consequências sejam minimizadas.

Considera-se que a previsão do absenteísmo, por meio da inteligência computacional, proposta neste trabalho, contribui no sentido de nortear quanto aos períodos mais adequados para prever o absenteísmo, de acordo com o objetivo de novas pesquisas doravante iniciadas.

A redução dos atributos, mediante a teoria dos *Rough Sets*, permite que novos estudos iniciem por meio dos atributos já identificados neste trabalho como relevantes, evitando um dispêndio de tempo com a coleta de dados não relevantes.

Desta forma, a análise aqui levantada abre caminho para novos trabalhos, com outras técnicas, que podem apresentar resultados melhores devido ao refinamento na escolha dos atributos relevantes que compõem o presente trabalho.

5.2 Contribuição para as empresas

No que se refere às empresas, a previsão do absenteísmo e a identificação de tendências absenteístas podem contribuir para a redução de: custos com pagamentos indevidos, abuso de licenças médicas, horas extras, excesso de colaborador, interrupções do fluxo de trabalho, substituição de empregados e contratação de trabalhadores temporários.

Além disso, a previsão e identificação dessas tendências promove um ambiente corporativo mais saudável e produtivo, melhora a relação de trabalho, bem como a estabilidade da empresa no mercado que está cada vez mais competitivo, e reduz gastos previdenciários.

5.3 Contribuição para os empregados

No caso dos empregados, a contribuição reside na possível melhoria da qualidade de vida no trabalho, reduzindo o contencioso entre os empregados e o estresse excessivo, que tendem a ser a causa de doenças que geram absenteísmo.

Além disso, a identificação de tendências absenteístas estimula as relações humanas dentro da empresa e contribui para a melhor distribuição do trabalho entre os empregados. Assim sendo, evita a sobrecarga de serviço, que pode gerar um ciclo vicioso de absenteísmo.

Concluindo, ao identificar o absenteísmo, é possível promover um bom clima organizacional, uma comunicação mais eficiente, feedbacks melhores do desempenho profissional, bem como a criação de políticas de reconhecimento e de metas individuais mais adequadas, que geram melhores resultados e maior satisfação ao empregado.

5.4 Contribuição teórica e limitação

A contribuição teórica para a academia e para o tema absenteísmo está no fato de este trabalho se mostrar como o pioneiro em aplicar técnicas de inteligência computacional no absenteísmo, colhendo resultados animadores. Este estudo, de certa forma, oferece um roteiro de aplicação das técnicas, e abre caminho para novos estudos sobre o assunto.

Indicadores bem definidos, para medir o absenteísmo, ajudam a controlar e diminuir o índice de absenteísmo nas empresas, pois podem ser comparados por períodos (mensal, bimestral, trimestral, quadrimestral, semestral, anual e bianual) por meio dos índices de

absenteísmo registrados. Dessa forma, é possível acompanhar as ações tomadas e os resultados atingidos de maneira mais objetiva, em relação ao aumento ou à redução do absenteísmo.

Considera-se como limitação deste trabalho a escassez de pesquisas sobre a aplicação de técnicas da inteligência computacional no estudo do absenteísmo, conforme comprovado na revisão bibliométrica aqui realizada, pois não foi possível fazer uma comparação dos resultados dos experimentos computacionais obtidos com outras pesquisas.

5.5 Contribuição da revisão bibliométrica

No decorrer da revisão bibliométrica, observou-se que existe uma diferença significativa no tratamento do absenteísmo, dependendo do país e da economia. Em geral, os países menos desenvolvidos mostram uma preocupação maior com a produtividade e o controle dos índices de absenteísmo, não levando em conta a qualidade de vida do empregado no trabalho.

Em países menos desenvolvidos e com mão de obra abundante, simplesmente não existem estudos relacionados ao absenteísmo, provavelmente pela facilidade de substituição de empregados com problemas de saúde e com baixa produtividade, além da inexistência de programas de qualidade de vida.

Na revisão bibliométrica, observou-se ainda que, no Brasil, de maneira geral, as empresas se preocupam muito mais com a produtividade e o controle dos índices de absenteísmo do que com a qualidade de vida do empregado, apesar de algumas empresas brasileiras apresentarem programas voltados para a qualidade de vida.

Nesse ponto, percebe-se que os estudos sobre o absenteísmo, as relações do trabalho no Brasil, e a qualidade de vida dos empregados, ainda são carentes de estudos. Tal carência gera um custo que, por sua vez, acaba reduzindo a competitividade das empresas, tanto no mercado interno quanto externo, em que os produtos e serviços são encarecidos pelo absenteísmo, seja pelo aumento do custo direto e indireto, pela redução da qualidade, ou ainda pelo aumento dos custos previdenciários.

Outro ponto observado no decorrer da revisão bibliométrica foi a abrangência da problemática que envolve o absenteísmo, conforme tratado por Miraglia e Johns (2018) em um modelo multinível da dinâmica social do fenômeno do absenteísmo, o que demonstra a complexidade do assunto em diversos níveis, desde o indivíduo até a sociedade e nação de origem, passando pelo grupo de trabalho e pela referência de vida, organização, ocupação e comunidade.

Yu et al. (2019) apresentaram um estudo recente sobre o estresse térmico, que afeta a saúde dos trabalhadores, reduz a produtividade do trabalho e causa absenteísmo. O estresse térmico transcende a sociedade e a nação de origem, acrescentando mais um fator no estudo das causas do absenteísmo. As mudanças nas condições climáticas sofridas no planeta elevam o estudo do absenteísmo a uma abrangência mundial.

Ainda na revisão bibliométrica, percebeu-se que, nos diversos estudos existentes, não há um consenso quanto aos atributos a serem utilizados, e os motivos do absenteísmo diferem significativamente. Alguns estudos, por exemplo, consideram férias, licença maternidade e abonos diversos amparados por lei como absenteísmo, e quantificam esses abonos juntamente com as licenças médicas de até 15 dias. Essa forma de quantificar o absenteísmo pode dificultar a tomada de decisão, uma vez que o absenteísmo gerenciável normalmente encontra-se nas licenças médicas de até 15 dias.

Percebeu-se também no decorrer do trabalho que quando acaba a relação ética entre a empresa e o empregado, o custo gerado, seja por absenteísmo ou presenteísmo aumenta e torna a empresa menos competitiva no mercado e mais vulnerável ao fracasso.

Por fim, na revisão bibliométrica, observou-se que uma das melhores práticas para reduzir o absenteísmo é estimular as relações humanas dentro do ambiente de trabalho, pois, como já foi mencionado, o absenteísmo apresenta causas psíquicas, sociais, afetivas, e não, necessariamente, materiais.

5.6 Contribuição para estudos futuros

Para estudos futuros, tem-se a possibilidade de aplicação de outras técnicas de inteligência computacional associadas às que foram aplicadas neste trabalho, como a lógica *Fuzzy*, na tentativa de reduzir os ruídos apresentados nos experimentos computacionais.

Outras técnicas de inteligência computacional ainda podem ser aplicadas para prever o absenteísmo, assim como para identificar tendências absenteístas. A base de dados, aqui utilizada, também permite a realização de outros experimentos com objetivos distintos, como associação, classificação, detecção de desvio, estimativa, regressão, agrupamento e sumarização.

Diante disso, está em consideração uma continuidade deste estudo, com as diversas técnicas de inteligência computacional associadas, também conhecidas como sistemas híbridos inteligentes.

O estudo aqui apresentado não teve a pretensão de saturar o assunto, pelo contrário, buscou-se contribuir para o entendimento do fenômeno absenteísmo, por meio da previsão e da identificação de tendências absenteístas, ao aplicar técnicas de inteligência computacional.

5.7 Memorial de Publicações

- Em Periódico Internacional:

FERREIRA, R. P.; MARTINIANO, A.; NAPOLITANO, D.; FARIAS, E. B. P.; SASSI, R. J. Artificial neural network and their application in the prediction of absenteeism at work. International Journal of recent scientific research, v. 8, p. 23332-23334, 2018. v. 8, p. 1-8, 2018. ISSN: 0976-3031.

- Em Periódico nacional:

FERREIRA, R. P.; MARTINIANO, A. S.; SASSI, R. J. Intelligent dashboard to support for decision making in business courier company. Revista Gestão & Tecnologia, v. 16, p. 39-72, 2016. ISSN: 1677-9479.

- Capítulo de livro:

MARTINIANO, A.; FERREIRA, R. P.; FARIAS, E. B. P.; GATTO, D. D. O.; SASSI, R. J. Mineração de dados aplicada ao Absenteísmo e a Ergonomia. In: Editora Poisson. (Org.). Tópicos em Administração - Capítulo 17 - Mineração de dados aplicada ao Absenteísmo e a Ergonomia. 1ª ed. Belo Horizonte: Poisson, v.8, p. 204-217, 2018.

5.8 Bases de Dados cedidas para UCI *Machine Learning Repository*

- Base de Dados cedida para UCI *Machine Learning Repository* do Centro de Aprendizado de Máquina e Sistemas Inteligentes da Universidade da Califórnia:

1. MARTINIANO, A.; FERREIRA, R. P.; SASSI, R. J. Absenteeism at work Data set - Base de Dados cedida para UCI *Machine Learning Repository* (<http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>). 2018.
2. FERREIRA, R. P.; MARTINIANO, A.; SASSI, R. J. Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set - Base de Dados cedida para UCI *Machine Learning Repository*. (<https://archive.ics.uci.edu/ml/datasets/Behavior+of+the+urban+traffic+of+the+city+of+Sao+Paulo+in+Brazil>). 2018.
3. FERREIRA, R. P.; MARTINIANO, A.; FERREIRA, A.; ROMERO, M.; SASSI, R. J. Container Crane Controller Data Set - Base de Dados cedida para UCI *Machine Learning Repository* (<http://archive.ics.uci.edu/ml/datasets/Container+Crane+Controller+Data+Set>). 2018.
4. FERREIRA, R. P.; MARTINIANO, A. S.; FERREIRA, A.; FERREIRA, A.; SASSI, R. J. Daily Demand Forecasting Orders Data Set - Base de Dados cedida para UCI *Machine Learning Repository* (<https://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders>). 2017.

As bases de dados foram cedidas para UCI *Machine Learning Repository* com o objetivo de popularizar os dados das pesquisas realizadas na Universidade Nove de Julho e para que outros estudantes e pesquisadores possam reproduzir os experimentos.

- Citações da base de dados *Absenteeism at work Data set* no Google Acadêmico:

- ARAUJO, V. S.; REZENDE, T. S.; GUIMARÃES, A. J.; ARAUJO, V. J. S.; CAMPOS SOUZA, P. V. A hybrid approach of intelligent systems to help predict absenteeism at work in companies. *SN Applied Sciences*, v.1, n.6, p.536, 2019. (Brasil).

- GAYATHRI, T. Data mining of Absentee data to increase productivity. *International Journal of Engineering and Techniques*. v.4, n.3, p.478-480, 2018. (Índia).
- RAJAB, S.; SHARMA, V. A review on the applications of neuro-fuzzy systems in business. *Artificial Intelligence Review*, v. 49, n. 4, p. 481-510, 2018. (Índia).
- RAJAB, S.; SHARMA, V. Neuro-fuzzy approach in business applications: a decade survey. *International Journal of Business Information Systems*, v. 26, n. 4, p. 451-487, 2017. (Índia).
- SCAFFIDI, C. The impact of human-centric design on the adoption of information systems: A case study of the spreadsheet. In: *Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on*. IEEE, 2016. p. 1-7. (Estados Unidos).
- ROSAS-DANIEL, J. A.; RODRÍGUEZ-ELIAS, O. M.; VELAZQUEZ-MENDOZA, M. D. J.; ROSE-GÓMEZ, C. E. A Literature Review on the Use of Soft Computing in Support of Human Resource Management. *Research in Computing Science*, 80, 107-117, 2014. (México).

- Citações da base de dados *Absenteeism at work Data set* em outra base:

- LARSSON, C. Case Representation Methodology for a Scalable Case-Based Reasoning. Malardalen University. School of Innovation, Design and Engineering - Vasteras, Sweden, 2018. (Suécia).

- Citações da base *Daily Demand Forecasting Orders Data set* no Google Acadêmico:

- QASEM, S. N. A PART-Based Algorithm for Prediction of Daily Demand Orders. *International Journal of Computer Science and Network Security - IJCSNS*, v.19, n.9, p.91-95, 2019. (Arábia Saudita e Iêmen).
- KERDPRASOP, N.; CHANSILP, K.; KERDPRASOP, K.; CHUAYBAMROONG, P. Anomaly Detection with Machine Learning Technique to Support Smart Logistics. In: *International Conference on Computational Science and Its Applications*. Springer, Cham, p. 461-472, 2019. (Tailândia).
- ALSANAD, A. Forecasting Daily Demand of Orders Using Random Forest Classifier. *International Journal of Computer Science and Network Security - IJCSNS*, v.18, n.4, p.79-83, 2018. (Arábia Saudita).

5.9 Registro de programa de computador oriundo da pesquisa

- FERREIRA, R. P.; MARTINIANO, A.; SASSI, R. J. Software Tool for Absenteeism Prediction. 2018. Patente: Programa de Computador. Número do registro: BR512018001297-3, data de registro: 14/01/2018, título: "*Software Tool for Absenteeism Prediction*", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial. (Anexo B).

REFERÊNCIAS BIBLIOGRÁFICAS

- ADDAE, H. M.; JOHNS, G.; BOIES, K. The legitimacy of absenteeism from work: a nine nation exploratory study. *Cross Cultural Management: An International Journal*, v. 20, n. 3, p. 402-428, 2013.
- AGAPITO, S. M.; DE SOUSA, F. C. A influência da satisfação profissional no absentismo laboral. *Revista portuguesa de saúde pública*, v. 28, n. 2, p. 132-139, 2010.
- AGUIAR, G. A. S.; OLIVEIRA, J. R. Absenteísmo: Suas principais causas e consequências em uma empresa do ramo de saúde. *Revista de Ciências Gerenciais*, vol. 13, n.18, p.95-113, 2009.
- AHN, S.; ANDERSON, K.; LEE, S. Exploration of the effect of workers' influence network on their absence behavior using agent-based modeling and simulation. In: *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*. IEEE Press, p. 3059-3065, 2013.
- AHN, S.; LEE, S. Development of model of workers mental processes related to absence norm as behavior rule in agent-based simulation. In: *Proceedings of the Winter Simulation Conference*. Winter Simulation Conference, p. 3484-3492, 2011.
- AKHTAR, R.; SANTOS, J. R. Risk analysis of hurricane disruptions on workforce and interdependent regional sectors. In: *Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, p. 41-46, 2013.
- ALMOND, J. M. The management of absenteeism. *The Production Engineer, IEEE*. v. 54, n. 12, p. 649-654, 1975.
- ANDRADE, R. D.; FERRARI JUNIOR, G. J.; CAPISTRANO, R., TEIXEIRA, C. S.; SILVA BELTRAME, T.; PEREIRA G. F. E. Absenteísmo na Indústria está associado com o Trabalho em Turnos e com Problemas no Sono. *Ciencia e trabalho*, v.19, n.58, p.35-41, 2017.
- ANUMALLA, K. Sistema de Gestão da Pré-Processamento de Dados. Dissertação de Mestrado em Ciência - Universidade de Akron, USA, 64 p., 2007.
- APPOLINÁRIO, F. Dicionário de Metodologia Científica. 2. ed. São Paulo: Atlas, 2011.
- ARDABILI, S. F.; NAJAFI, B.; SHAMSHIRBAND, S.; MINAEI BIDGOLI, B.; DEO, R. C.; CHAU, K. W. Computational intelligence approach for modeling hydrogen production: A review. *Engineering Applications of Computational Fluid Mechanics*, v.12, n.1, p.438-458, 2018.
- BACHOUCH, R. B.; GUINET, A.; HAJRI-GABOUJ, S. An optimization model for task assignment in home health care. In: *Health Care Management (WHCM), 2010 IEEE Workshop on*. IEEE, p. 1-6, 2010.
- BAGGA, S.; SINGH, G. N. Conceptual Three Phase Iterative Model of KDD. *International Journal of Computers & Technology*, v.2, n.1, p.6-8, 2012.
- BAKKER, A. B.; DEMEROUTI, E. The job demands-resources model: State of the art. *Journal of managerial psychology*, v.22, n.3, p.309-328, 2007.

- BAMBERGER, P.; BIRON, M. Group norms and excessive absenteeism: the role of peer referent others. *Organizational Behavior and Human Decision Processes*, v.103, p.179-196, 2007.
- BANSBACK, N.; ZHANG, W.; WALSH, D.; KIELY, P.; WILLIAMS, R.; GUH, D.; ANIS, A.; YOUNG, A. Factors associated with absenteeism, presenteeism and activity impairment in patients in the first years of RA. *Rheumatology*, 51(2), 375-384, 2012.
- BECKER, H. S. *Tricks of the trade: How to think about your research while you're doing it*. University of Chicago press, 2008.
- BENNEDSEN, M.; TSOUTSOURA, M.; WOLFENZON, D. Drivers of Effort: Evidence from Employee Absenteeism. *Journal of Financial Economics*, p1-53, 2018.
- BERGSTRÖM, G.; HAGBERG, J.; BUSCH, H.; JENSEN, I.; BJÖRKLUND, C. Prediction of sickness absenteeism, disability pension and sickness presenteeism among employees with back pain. *Journal of occupational rehabilitation*, 24(2), 278-286, 2014.
- BERKHIN, P. A survey of clustering data mining techniques. In: *Grouping multidimensional data*. Springer, Berlin, Heidelberg. p.25-71, 2006.
- BHUI, K. S.; DINOS, S.; STANSFELD, S. A.; WHITE, P. D. A synthesis of the evidence for managing stress at work: a review of the reviews reporting on anxiety, depression, and absenteeism. *Journal of environmental and public health*, v.2012, p.1-21, 2012.
- BIGUS, J. P. *Data Mining with Neural Network: Solving Business Problems from Applications Development to Decision Support*. Mcgraw-Hill, 1996.
- BIRON, M.; BAMBERGER, P. Aversive Workplace Conditions and Absenteeism: Taking Referent Group Norms and Supervisor Support Into Account. *Journal of Applied Psychology*, v.97, n.4, p.901-912, 2012.
- BÖCKERMAN, P.; LAUKKANEN, E. Predictors of sickness absence and presenteeism: does the pattern differ by a respondent's health? *Journal of Occupational and Environmental Medicine*, v. 52, n. 3, p. 332-335, 2010.
- BÖCKERMAN, P.; LAUKKANEN, E. Presenteeism in Finland: Determinants by gender and the sector of economy. *Ege Academic Review*, v. 9, n. 3, 2009.
- BOND, A. H.; GASSER, L. (Ed.). *Readings in distributed artificial intelligence*. Morgan Kaufmann, 2014.
- BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. *Redes Neurais Artificiais: Teoria e Aplicações (2 ed.)*. Rio de Janeiro: LTC, 2011.
- BREARD, G. T. *Evaluating self-organizing map quality measures as convergence criteria*. Master thesis, University of Rhode Island, 2017.
- BRAHAMI, M.; ATMANI, B.; MATTA, N. Dynamic knowledge mapping guided by data mining: application on healthcare. *Journal of Information Processing Systems*, v.9, n.1, p.1-30, 2013.
- BRATSBERG, B.; FEVANG, E.; RØED, K. Job loss and disability insurance. *Labour Economics*, 24, p.137-150, 2013.

- BRITT, T. W.; JEX, S. M. Thriving under stress: Harnessing demands in the workplace. Oxford University Press, USA, 2015.
- BROWN, H. E.; GILSON, N. D.; BURTON, N. W.; BROWN, W. J. Does physical activity impact on presenteeism and other indicators of workplace well-being? *Sports Medicine*, 41(3), 249-262, 2011.
- BUSTILLOS, A. S.; VARGAS III. K. G.; GOMERO-CUADRA. R. Work productivity among adults with varied Body Mass Index: Results from a Canadian population-based survey. *Journal of Epidemiology and Global Health*, 5, p.191-199, 2015.
- CAMILO, C. O.; SILVA, J. C. Da. Mineração de dados: Conceitos, tarefas, técnicas e ferramentas. Technical Report. Instituto de Informática da Universidade Federal de Goiás, 2009.
- CAMPOS, M. M.; SAITO, K. Sistemas inteligentes em controle e automação de processos. *Ciência Moderna*, 2004.
- CASTLE, N. G.; FERGUSON-ROME, J. C. Influence of nurse aide absenteeism on nursing home quality. *The Gerontologist*, v. 55, n. 4, p. 605-615, 2014.
- CHAN, W. H. Labour - Turnover and Absenteeism. *The Production Engineer*, IEEE, p.202-210, 1968.
- CICEI, C. C.; MOHOREA, L.; TEODORU, A. A. Investigating two predictors of sickness presenteeism on a romanian sample. The case of performance-based self-esteem and overcommitment. *Procedia-Social and Behavioral Sciences*, v. 78, p. 325-329, 2013.
- COCKER, F.; MARTIN, A.; SCOTT, J.; SANDERSON, K. Depression-related absenteeism and presenteeism amongst employed Australian adults: Do working conditions influence the health and economic consequences? *Journal of Affective Disorders*, 122, S34, 2010.
- COCKER, F.; MARTIN, A.; SCOTT, J.; VENN, A.; OTAHAL, P.; SANDERSON, K. Factors associated with presenteeism among employed Australian adults reporting lifetime major depression with 12-month symptoms. *Journal of Affective Disorders*, 135(1), p.231-240, 2011.
- COHEN, A.; GOLAN, R. Predicting absenteeism and turnover intentions by past absenteeism and work attitudes: An empirical examination of female employees in long term nursing care facilities. *Career Development International*, v.12, n.5, p.416-432, 2007.
- COLLINS, P. An exploratory study into adolescent mental health difficulties in the United Kingdom in the 21st century. Tese de Doutorado. Cardiff Metropolitan University, 57f. 2017.
- CORREIOS. Soluções que aproximam. Disponível em: <<http://www.correios.com.br/>>. Acesso em: 29 out. 2019.
- CRUZ, C. R. B.; SHIRASSU, M. M.; BARBOSA, P. L. Q.; SANTANA, A. M. R. D. Mental disorders as a cause of absenteeism among public workers in São Paulo. *Archives of Clinical Psychiatry (São Paulo)*, 38(6), p.265-266, 2011.
- CUCCHIELLA, F.; GASTALDI, M.; RANIERI, L. Managing absenteeism in the workplace: the case of an Italian multiutility company. *Procedia-Social and Behavioral Sciences*, v.150, p. 1157-1166, 2014.

- CUNHA, J. B.; BLANK, V. L. G.; BOING, A. F. Tendência Temporal de Afastamento do Trabalho em Servidores Públicos (1995-2005). *Revista Brasileira de Epidemiologia*. vol.12 n. 2 São Paulo, 2009.
- COSTA, J. A. F.; ANDRADE NETTO, M. L. Segmentação de Mapas Auto-Organizáveis com Espaço de Saída 3-D. *Revista Controle & Automação*.v.18, n.2, 2007.
- DARR, W.; JOHNS, G. Work strain, health, and absenteeism: a meta-analysis. *Journal of occupational health psychology*, v. 13, n. 4, p. 293, 2008.
- DE LIMA SANTANA, L.; SARQUIS, L. M. M.; BREY, C.; MIRANDA, F. M. D.; FELLI, V. E. A. Absenteísmo por transtornos mentais em trabalhadores de saúde em um hospital no sul do Brasil. *Revista Gaúcha de Enfermagem*, 37(1), 2016.
- DEMEROUTI, E.; LE BLANC, P. M.; BAKKER, A. B.; SCHAUFELI, W. B.; HOX, J. Present but sick: a three-wave study on job demands, presenteeism and burnout. *Career Development International*, 14(1), p.50-68, 2009.
- DEMIR, C.; CERGIPOZAN, R. Determinants of Patent Protection Regimes: A Self-Organizing Map Approach. *Review of Economic Perspectives*, v.18, n.3, p.261-283, 2018.
- DESPIÉGEL, N.; DANCHENKO, N.; FRANÇOIS, C.; LENSBERG, B.; DRUMMOND, M. F. The use and performance of productivity scales to evaluate presenteeism in mood disorders. *Value in Health*, 15(8), p.1148-1161, 2012.
- DEW, K.; KEEFE, V.; SMALL, K. Choosing to work when sick: workplace presenteeism. *Social science & medicine*, v. 60, n. 10, p.2273-2282, 2005.
- DEZA, M. M.; DEZA, E. Encyclopedia of distances. In: *Encyclopedia of Distances* (4.ed.) Springer, Berlin, Heidelberg, 2016.
- DHIMAN, G.; KAUR, A. A hybrid algorithm based on particle swarm and spotted hyena optimizer for global optimization. In *Soft Computing for Problem Solving*. Springer, Singapore, p.599-615, 2019.
- FALLAH, S.; DEO, R.; SHOJAFAR, M.; CONTI, M.; SHAMSHIRBAND, S. Computational intelligence approaches for energy load forecasting in smart energy management grids: state of the art, future challenges, and research directions. *Energies*, v.11, n.3, 596, p.1-31, 2018.
- FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMITH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, v.39, p.27-34, 1996.
- FERNANDES, M. A.; SANTOS, J. D. M.; MORAES, L. M. V. D.; LIMA, J. S. R.; FEITOSA, C. D. A.; SOUSA, L. F. C. Mental and behavioral disorders in workers: a study on work leave. *Revista da Escola de Enfermagem da USP*, v.52, p.1-8, 2018.
- FERREIRA, L. A. L; FERREIRA, L. L. Depressão no trabalho da enfermagem: revisão de literatura. *Universitas: Ciências da Saúde*, v. 13, n.1, p. 41-48, 2015.
- FIGLIOLI, J. O. *Psicologia para administradores: integrando teoria e prática* (8 ed.). São Paulo: Atlas, 2013.
- FOTOVATIKHAH, F.; HERRERA, M.; SHAMSHIRBAND, S.; CHAU, K. W.; ARDABILI, S. F.; PIRAN, M. J. Survey of computational intelligence as basis to big flood management:

Challenges, research directions and future work. *Engineering Applications of Computational Fluid Mechanics*, v.12, n.1, p.411-437, 2018.

FROOMAN, J.; MENDELSON, M. B.; KEVIN, M. J. Transformational and passive avoidant leadership as determinants of absenteeism. *Leadership & Organization Development Journal*, v. 33, n. 5, p.447-463, 2012.

GACHUI, J. G.; WERE, S. Investigating the Causes of Presenteeism Among Secondary School Employees in Nairobi County. *International Journal of Social Sciences Management and Entrepreneurship*, v.1, n.1, p.18-35, 2014.

GARCÍA, A. Problemática multifactorial del absentismo laboral, el presentismo y la procrastinación en las estructuras en que se desenvuelve el trabajador. *Medicina y Seguridad del Trabajo*, 57(223), p.111-120, 2011.

GIL, A. C. *Como Elaborar Projetos de Pesquisa* (5ª. ed.). São Paulo: Atlas, 2010.

GISBERT, J. P.; COOPER, A.; KARAGIANNIS, D.; HATLEBAKK, J.; AGRÉUS, L.; JABLONOWSKI, H.; NUEVO, J. Impact of gastroesophageal reflux disease on work absenteeism, presenteeism and productivity in daily life: a European observational study. *Health and quality of life outcomes*, 7(1), 90, 2009.

GODØY, A.; DALE-OLSEN, H. Spillovers from gatekeeping—Peer effects in absenteeism. *Journal of Public Economics*, v.167, p.190-204, 2018.

GOLDSCHMIDT, R. R.; PASSOS, E.; BEZERRA E. *Data Mining: conceitos, técnicas, algoritmos, orientações e aplicações*. (2 ed.). Campus: Rio de Janeiro, 2015.

GOLDSCHMIDT, R. R. *Uma Introdução à Inteligência Computacional: fundamentos, ferramentas e aplicações*. Rio de Janeiro. Brasil: IST-Rio, 2010.

GUTIÉRREZ, F. R.; BALDEÓN, J. P.; BELTRÁN, R. Z.; YESÁN, C. L. Prevalencia, impacto en la productividad y costos totales de las principales enfermedades en los trabajadores de un hospital al sur del Perú en el año 2003. *Rev Med Hered* 17 (1), 2006.

HAEFFNER, R.; KALINKE, L. P.; FELLI, V. E. A.; MANTOVANI, M. D. F.; CONSONNI, D.; SARQUIS, L. M. M. Absenteísmo por distúrbios musculoesqueléticos em trabalhadores do Brasil: milhares de dias de trabalho perdidos. *Revista Brasileira de Epidemiologia*, v.21, p.1-11, 2018.

Haidar, V. La aparición y declinación del ausentismo como problema para el gobierno de la población trabajadora (Argentina, siglo XX). *Trabajo y sociedad*, n. 20, p.395-425, 2013.

HALBESLEBEN, J. R. B.; WHITMAN, M. V.; CRAWFORD, W. S. A dialectical theory of the decision to go to work: Bringing together absenteeism and presenteeism. *Human Resource Management Review*, 24, p.177-192, 2014.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. (3 ed.) Waltham, MA: Morgan Kaufmann, 2011.

HARGRAVE, G. E.; HIATT, D.; ALEXANDER, R.; SHAFFER, I. A. EAP treatment impact on presenteeism and absenteeism: Implications for return on investment. *Journal of Workplace Behavioral Health*, v.23, n.3, p.283-293, 2008.

- HAYKIN, S. *Neural networks and learning machines* (3.ed). Upper Saddle River, NJ, USA: Pearson, 2009.
- HAYKIN, S. *Redes Neurais - Princípios e Práticas*. Bookman. (2 ed.) Porto Alegre, 2001.
- HEBB, D. O. *The Organization of Behavior: Neuropsychological Theory*, N. Y.:Willey, 1949.
- HEIDARI, M.; SHAMSI, H. Analog programmable neuron and case study on VLSI implementation of Multi-Layer Perceptron (MLP). *Microelectronics Journal*, p.1-24, 2018.
- HEMP, P. Presenteeism: At Work - but out of it. *Business Review Journal*, 82(10), p.49-58, 2004.
- HENSING, G.; BERTILSSON, M.; AHLBORG, G.; WAERN, M.; VAEZ, M. Self-assessed mental health problems and work capacity as determinants of return to work: a prospective general population-based study of individuals with all-cause sickness absence. *BMC psychiatry*, 13(1), 259, 2013.
- HERNÆS, Ø. Activation against absenteeism—Evidence from a sickness insurance reform in Norway. *Journal of health economics*, v.62, p.60-68, 2018.
- HITO, S. C. *Condições de Trabalho e Absenteísmo Odontológico em uma Indústria Frigorífica no Brasil: uma contribuição à saúde Ocupacional (Dissertação de Mestrado)*. Centro Universário Senac. São Paulo, SP, 2007.
- HUTTING, N.; ENGELS, J. A.; HEERKENS, Y. F.; STAAL, J. B; NIJHUIS-VAN DER SANDEN, M. W. Development and measurement properties of the Dutch version of the Stanford Presenteeism Scale (SPS-6). *Journal of occupational rehabilitation*, 24(2), 268-277, 2014.
- INOUE, K. C.; MATSUDA, L. M.; SILVA, D. M. P. P.; UCHIMURA, T. T.; MATHIAS, T. A. F. Absenteísmo-doença da equipe de enfermagem em unidade de terapia intensiva. *Revista Brasileira de Enfermagem*, v.61, n.2, p.209-214, 2008.
- ISETTI, D.; MEYER, T. Workplace productivity and voice disorders: A cognitive interviewing study on presenteeism in individuals with spasmodic dysphonia. *Journal of Voice*, v. 28, n. 6, p. 700-710, 2014.
- JANTAN, H.; HAMDAN, A.R.; OTHMAN, Z.A. Human Talent Prediction in HRM using C4.5 Classification Algorithm. *International Journal on Computer Science and Engineering*, 2(8), p.2526-2534, 2010.
- JANTAN, H.; HAMDAN, A.R.; OTHMAN, Z.A. Towards applying Data Mining Techniques for Talent Mangement. In *International Conference on Computer Engineering and Applications*, (2). IACSIT, 2009.
- JANTAN, H.; HAMDAN, A.R; OTHMAN, Z.A. Potential intelligent techniques in Human Resource Decision Support System (HR DSS). In *International Symposium on Information Technology*, v.3, p.1-9, 2008.
- JARSKE, J. M.; SEABRA, A. G.; SILVA, L. A. Self-Organizing Maps Applied as Analysis Tool of Reading Cognitive Test. *IEEE Latin America Transactions*, v.16, n.6, p.1817-1824, 2018.

- JODAS, D. A.; HADDAD, M. C. L. Síndrome de Burnout em Trabalhadores de Enfermagem de um Pronto Socorro de Hospital Universitário. *Acta Paulista de Enfermagem*, v.22, n.2, p.192-197, 2009.
- JOHNS, G. Absenteeism and presenteeism: Not at work or not working well. In C. L. Cooper, & J. Barling (Eds.), *The Sage handbook of organizational behavior*, v.1, p.160-177, London, 2008.
- JOHNS, G. Absenteeism. *Wiley Encyclopedia of Management*, p. 1-3, 2015.
- JOHNS, G. Attendance dynamics at work: the antecedents and correlates of presenteeism, absenteeism, and productivity loss. *Journal of Occupational Health Psychology*, v.16, n.4, p. 483, 2011.
- JOHNS, G. Attitudinal and nonattitudinal predictors of two forms of absence from work. *Organizational behavior and human performance*, v. 22, n. 3, p. 431-444, 1978.
- JOHNS, G. How methodological diversity has improved our understanding of absenteeism from work. *Human resource management review*, v. 13, n. 2, p. 157-184, 2003.
- JOHNS, G. Substantive and methodological constraints on behavior and attitudes in organizational research. *Organizational behavior and human decision processes*, v. 49, n. 1, p. 80-104, 1991.
- JOHNS, G. The essential impact of context on organizational behavior. *Academy of management review*, v.31, n.2, p.386-408, 2006.
- JOHNS, G. Presenteeism in the workplace: A review and research agenda. *Journal of Organizational Behavior*, v. 31, n. 4, p. 519-542, 2010.
- JONES, D. A.; NOZICK, L. K.; TURNQUIST, M. A.; SAWAYA, W. J. Pandemic influenza, worker absenteeism and impacts on freight transportation. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (p. 206-206). IEEE, january, 2008.
- JUNOR, A.; O'BRIEN, J.; O'DONNELL, M. Welfare wars: public service frontline absenteeism as collective resistance. *Qualitative Research in Accounting & Management*, v. 6, n. 1/2, p. 26-40, 2009.
- JOTHI, N.; RASHID, N. A. A.; e HUSAIN, W. Data mining in healthcare—a review. *Procedia Computer Science*, v.72, p.306-313, 2015.
- JOURDAIN, G; CHÊNEVERT, D. The moderating influence of perceived organizational values on the burnout-absenteeism relationship. *Journal of Business and Psychology*, v.30, n.1 p.77-191, 2015.
- KARANIKA-MURRAY, M.; PONTES. H. M.; GRIFFITHS, M. D.; BIRON, C. Sickness presenteeism determines job satisfaction via affective-motivational states. *Social Science & Medicine*, v.139, p.100-106, 2015.
- KAPANOVA, K. G.; DIMOV, I.; SELIER, J. M. A neural network sensitivity analysis in the presence of random fluctuations. *Neurocomputing*, v.224, p.177-183, 2016.
- KASKI, S.; KOHONEN, T. Winner-Takes-All Networks. *Triennial Report 1994 – 1996, Neural Networks Research Centre & Laboratory of Computer and Information Science, Helsinki University of Technology, Finland*, p. 72-75, 1997.

- KIM, M. S.; PARK, J. B.; MIN, K. B.; LEE, K. J.; KWON, K. Presenteeism among self-employed workers: Korean working conditions survey. *Annals of occupational and environmental medicine*, v.26, n.1, p.32, 2014.
- KIVILUOTO, K. Topology preservation in self-organizing maps. in *Proc. International Conference Neural Networks*, Washington, DC, p. 294-299, 1996.
- KNANI, M. Exploratory study of the impacts of new technology implementation on burnout and presenteeism. *International Journal of Business and Management*, v. 8, n. 22, p. 92, 2013.
- KNIES, S.; BOONEN, A.; SEVERENS, J. L. Do the Washington Panel recommendations hold for Europe: investigating the relation between quality of life versus work-status, absenteeism and presenteeism. *Cost Effectiveness and Resource Allocation*, v. 12, n. 1, p.12-24, 2014.
- KOHONEN, T. Essentials of the self-organizing map. *Neural networks*, v. 37, p. 52-65, 2013.
- KOHONEN, T. *Self-Organizing Maps*. New York (3 ed.). Springer, 2001.
- KOMOROWSKI, J.; PAWLAK, Z.; POLKOWSKI, L.; SKOWRON, A. Rough sets: A tutorial. *Rough fuzzy hybridization: A new trend in decision-making*, p.3-98, 1999.
- KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, v.160, p.3-24, 2007.
- KOTTWITZ, M. U.; SCHADE, V.; BURGER, C.; RADLINGER, L.; ELFERING, A. Time Pressure, Time Autonomy, and Sickness Absenteeism in Hospital Employees: A Longitudinal Study on Organizational Absenteeism Records. *Safety and Health at Work*, v.9, n.1, p.109-114, 2018.
- KRISTENSEN, K.; JUHL, H.J.; ESKILDSEN, J.K.; NIELSEN, J.; FREDERIKSEN, N.; BISGAARD, C. Determinants of absenteeism in a large Danish bank, *International Journal of Human Resource Management*, v.17, p.1645-1658, 2006.
- LAAKSONEN, M.; PITKANIEMI, J.; RAHKONEN, O.; LAHELMA E. Work arrangements, physical working conditions, and psychosocial working conditions as risk factors for sickness absence: Bayesian analysis of prospective data. *Annals of Epidemiology*, v.20, n.5, p.332-338, 2010.
- LAHTI, J.; LAAKSONEN, M.; LAHELMA, E.; RAHKONEN, O. The impact of physical activity on sickness absence. *Scandinavian journal of medicine & science in sports*, v.20, n.2, p.191-199, 2010.
- LAINE, S. Selecting the variables that train a Self-Organizing Map (SOM) which best separates predefined clusters. *International Conference on Neural Information Processing - ICONIP 2002*, Singapore, p.1961-1966, 2002.
- LANDRY, M.; MILLER, C. Presenteeism: are we hurting the patients we are trying to help? *Journal of General Internal Medicine*, v.25, n.11, p.1142-1143, november, 2010.
- LARDON, A.; DUBOIS, J. D.; CANTIN, V.; PICHÉ, M.; DESCARREAUX, M. Predictors of disability and absenteeism in workers with non-specific low back pain: A longitudinal 15-month study. *Applied ergonomics*, v.68, p.176-185, 2018.

- LAST, M.; KLEIN, Y.; KANDEL, A. Knowledge discovery in time series databases. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, v.31, n.1, p.160-169, 2001.
- LEÃO, A. L. D. M.; BARBOSA-BRANCO, A.; RASSI NETO, E.; RIBEIRO, C. A. N.; TURCHI, M. D. Absenteísmo-doença no serviço público municipal de Goiânia. *Revista Brasileira de Epidemiologia*, v.18, p.262-277, 2015.
- LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A.; CHENG, T. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, v.115, p.119-133, 2016.
- LI, W.; XU, S.; MENG, W. A Risk Prediction Model of Construction Enterprise Human Resources Based on Support Vector Machine. In *Second International Conference on Intelligent Computation Technology and Automation (1)*, p.945-948, 2009.
- LIEKE, L.; JOHNS, G.; LYONS, B. J.; TER HOEVEN, C. L. Why and when do employees imitate the absenteeism of co-workers? *Organizational Behavior and Human Decision Processes*, 134, 16-30, 2016.
- LINOFF, G. S.; BERRY, M. J. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 3^a ed., 2011.
- LOHAUS, D.; HABERMANN, W. Presenteeism: A review and research directions. *Human Resource Management Review*, v.29, n.1, p.43-58, 2019.
- LÖTTTERS, F.; BURDORF A. Prognostic factors for duration of sickness absence due to musculoskeletal disorders. *The Clinical journal of pain*, v.22, n.2, p.212-221, 2006.
- LOPES, K. M. O.; GARRIDO, G.; MENDONÇA, A. V.; SILVEIRA, M. A. Uma Abordagem Fuzzy para o estudo do Presenteísmo. *Revista Pensamento Contemporâneo em Administração*, v.11, n.2, p.136-150, 2017.
- LOZANO, J.; GÓMEZ, A.; PINO, R.; PUENTE, J.; PONTE, B. Lower bounds for estimating workforce size in a 24/7 company. In *Industrial Engineering and Engineering Management (IEEM)*, IEEE International Conference on (p.1675-1678). IEEE, december, 2013.
- LU, L.; COOPER, C. L.; LIN, H.Y. A cross-cultural examination of presenteeism and supervisory support. *Career Development International*, v.18, n.5, p.40- 56, 2013.
- LUND, T.; LABRIOLA, M.; CHRISTENSEN, K. B.; BULTMANN, U.; VILLADSEN, E. Physical work environment risk factors for long term sickness absence: prospective findings among a cohort of 5357 employees in Denmark. *BMJ*. v.332, n.7539, p.449-452, 2006.
- MACGREGOR, J. N.; CUNNINGHAM, B.; CAVERLEY, N. Factors in absenteeism and presenteeism: life events and health events. *Management Research News*, v.31, n.8, p.607-615, 2008.
- MARKUSSEN, S.; MYKLETUN, A.; RØED, K. The case for presenteeism—Evidence from Norway's sickness insurance program. *Journal of Public Economics*, v. 96, n. 11-12, p.959-972, 2012.

- MARQUES, D. D. O.; PEREIRA, M. S.; SOUZA, A. C. S.; VILA, V. D. S. C.; ALMEIDA, C. C. O. D. F.; OLIVEIRA, E. C. D. O absenteísmo-doença da equipe de enfermagem de um hospital universitário. *Revista Brasileira de Enfermagem*, v. 68, n. 5, p. 876-882, 2015.
- MARTÍNEZ-LÓPEZ, E.; SALDARRIAGA-FRANCO, J. Inactividad física y ausentismo en el ámbito laboral. *Revista de salud pública*, v. 10, p. 227-238, 2008.
- MARTINIANO, A.; FERREIRA, R. P.; SASSI, R. J.; AFFONSO, C. Application of a neuro fuzzy network in prediction of absenteeism at work. In *Information Systems and Technologies (CISTI)*, 7th Iberian Conference on (p. 1-4). IEEE, june, 2012.
- MATTAR, F. N. *Pesquisa de marketing: metodologia, planejamento*. 7.ed. Rio de Janeiro: Campus, 2014.
- McEWEN, B. S. Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiological reviews*, v. 87, n. 3, p. 873-904, 2007.
- McGRANDLE, J.; OHEMENG, F. L. The conundrum of absenteeism in the Canadian public service: A wicked problem perspective. *Canadian Public Administration*, v.60, n.2, p.215-240, 2017.
- MBP. Multiple Back-Propagation. Disponível em: <<http://mbp.sourceforge.net/>>. Acesso em: 29 out. 2019.
- MEI, Y. The Forecasting of Human Resource Supply Based on RBF Neural Network and Markov Transfer Matrix. In *Second International Symposium on Computational Intelligence and Design*, (1), p.385-388, 2009.
- MESA, F. R.; KAEMPFER, A. M. 30 años de estudio sobre ausentismo laboral en Chile: una perspectiva por tipos de empresas. *Revista médica de Chile*, v. 132, n. 9, p. 1100-1108, 2004.
- MESTRE, L. R.; FERREIRA, L. P. O impacto da disfonia em professores, queixas vocais, procura por tratamento, comportamento, conhecimento sobre cuidados com a voz e absenteísmo. *Revista da Sociedade Brasileira de Fonoaudiologia*, v. 16, n. 2, p. 240-241, 2011.
- MIKKULAINEN, R.; LIANG, J.; MEYERSON, E.; RAWAL, A.; FINK, D.; FRANCON, O.; RAJU, B.; SHAHRZAD, H.; NAVRUZYAN, A.; DUFFY, N.; HODJAT, B. Evolving deep neural networks. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. Academic Press. p.293-312, 2019.
- MILLER, E. H. Hampshire Sub-Centre: Chairman's address." Co-operation between the electrical and medical professions, and possible effect on absenteeism in industrial undertakings". *Journal of the Institution of Electrical Engineers-Part I: General*, IEEE, v. 93, n. 63, p. 129-133, 1945.
- MILLS, R. Predicting failure to return to work. *Internal medicine journal*, v. 42, n. 8, p. 924-927, 2012.
- MIRAGLIA, M.; JOHNS, G. A Multi-Level Integration of the Antecedents and Dynamics of Absenteeism. In: *Academy of Management Proceedings*. Briarcliff Manor, NY 10510: Academy of Management, p.1-40, 2018.
- MIRAGLIA, M.; JOHNS, G. Going to work ill: A meta-analysis of the correlates of presenteeism and a dual-path model. *Journal of Occupational Health Psychology*, v. 21, n. 3, p. 261, 2016.

- MIRANDA, G. J.; ARAUJO, T. S.; MARCELINO, I. A. O absenteísmo acadêmico e suas consequências mais óbvias. *Revista Gestão Universitária na América Latina-GUAL*, v. 10, n. 1, p. 172-189, 2017.
- MOREAU, M.; VALENTE, F.; MAK, R.; PELFRENE, E.; SMET, P.; DE BACKER, G.; KORNITZER, M. Obesity, body fat distribution and incidence of sick leave in the Belgian workforce: the Belstress study. *International journal of obesity*, v.28, n.4, p.574, 2004.
- NASCIMENTO JUNIOR, C. L.; YONEYAMA, T. *Inteligência Artificial em controle e automação*. São Paulo: Blucher/FAPESP, 2004.
- NING, C. The Application of Neural Network to the Allocation of Enterprise Human Resources. In *2nd International Conference on e-Business and Information System Security*. EBISS, p.1-4, 2010.
- NTNU. Faculty of Information Technology and Electrical Engineering. Department of Computer Science. Disponível em: <<https://www.ntnu.edu/idi/>>. Acesso em: 29 out. 2019.
- NYATHI, M.; JOOSTE, K. Working conditions that contribute to absenteeism among nurses in a provincial hospital in the Limpopo Province. *Curationis*, v. 31, n. 1, p. 28-37, 2008.
- OENNING, N. S. X.; CARVALHO, F. M.; LIMA, V. M. C. Fatores de risco para absenteísmo com licença médica em trabalhadores da indústria de petróleo. *Revista de Saúde Pública*, v. 48, p. 103-122, 2014.
- ONIKOYI, I. A.; AWOLUSI, O. D.; AYODEJI, B. M. Effect of Absenteeism on Corporate Performance: A Case Study of Cadbury Nigeria Plc, Ikeja, Lagos State, Nigeria. *British Journal of Marketing Studies*, v. 3, n. 2, p. 58-71, 2015.
- PARDO, J. M. V. La gestión de la incapacidad laboral algo más que una cuestión económica. *Medicina y Seguridad del Trabajo*, v. 64, n. 251, p. 131-160, 2018.
- PARKER, S. K.; MORGESON, F. P.; JOHNS, G. One hundred years of work design research: Looking back and looking forward. *Journal of applied psychology*, v. 102, n. 3, p. 403, 2017.
- PARKS, K. M.; STEELMAN, L. A. A. Organizational wellness programs: a meta-analysis. *Journal of occupational health psychology*, v. 13, n. 1, p. 58, 2008.
- PATTON, E.; JOHNS, G. Context and the social representation of absenteeism: Absence in the popular press and in academic research. *Human Relations*, v. 65, n.2, p. 217-240, 2012.
- PATTON, E.; JOHNS, G. Women's absenteeism in the popular press: evidence for a gender-specific absence culture, *Human Relations*, v.60, p.1579-1612, 2007.
- PAWLAK, Z. Rough sets. *International journal of computer & information sciences*, v.11, n.5, p.341-356, 1982.
- PAWLAK, Z.; SKOWRON, A. Rudiments of rough sets. *Information Sciences International journal*, v. 177, n. 1, p. 3-27, 2007.
- PHRIDVIRAJ, M. S. B.; GURURAO, C. V. Data mining—past, present and future—a typical survey on data streams. *Procedia Technology*, v.12, p.255-263, 2014.
- PIATETSKY-SHAPIRO, G.; MATHEUS, C. J.; CHAN, P. K. Systems for Knowledge Discovery in Data bases. *IEEE Transactions on Knowledge and Data Engineering*, 5, p.903-912, 1993.

- PICHLER, S.; ZIEBARTH, N. R. The pros and cons of sick pay schemes: Testing for contagious presenteeism and noncontagious absenteeism behavior. *Journal of Public Economics*, v.156, p. 14-33, 2017.
- POMBO, N.; ARAÚJO, P.; VIANA, J. Knowledge discovery in clinical decision support systems for pain management: A systematic review. *Artificial Intelligence in Medicine*, v.60, p.1-11, 2014.
- PRATER, T.; SMITH, K. Underlying factors contributing to presenteeism and absenteeism. *Journal of Business & Economics Research*, v.9, n.6, p.1-14, 2011.
- RANTANEN, I.; TUOMINEN, R. Relative magnitude of presenteeism and absenteeism and work-related factors affecting them among health care professionals. *International archives of occupational and environmental health*, v. 84, n. 2, p. 225-230, 2011.
- REBOLLAR, A. M.; GOMEZ, W. A. S.; GARCIA, R. A. B.; BALDIZON, Y. G.; LAVALLE, M. M.; HERNANDEZ, J. O. A Job Stress Predictive Model Evaluation Through Classifier's Algorithms. *IEEE Latin America Transactions*, v.16, n.1, p.178-185, 2018.
- REYNOLDS, A. C.; APPLETON, S. L.; GILL, T. K.; TAYLOR, A. W.; MCEVOY, R. D.; FERGUSON, S. A.; ADAMS, R. J. Sickness absenteeism is associated with sleep problems independent of sleep disorders: results of the 2016 Sleep Health Foundation national survey. *Sleep health*, v.3, n.5, p.357-361, 2017.
- RISTOSKI, P.; PAULHEIM, H. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, v. 36, p.1-22, 2016.
- RODRIGUES, M. L.; LIMENA, M. M. C. (Orgs.). *Metodologias multidimensionais em Ciências Humanas*. Brasília: Líber Livros Editora, 2006.
- RODRIGUES, T.; OLIVEIRA, M.D.; SA, A. B. Multicriteria model to allocate human resources in community care programmes. In *1st Portuguese Meeting in Bioengineering*, p.1-6, 2011.
- ROMERO, T. A.; STROM, K. J. Absenteeism in a Represented Environment. *International Journal of Humanities and Social Science*, v. 1, n.15, p.1-11, 2011.
- ROSAS-DANIEL, J.A.; RODRÍGUEZ-ELIAS, O.M.; VELAZQUEZ-MENDOZA M. DE J.; ROSE-GÓMEZ C. E. A Literature Review on the Use of Soft Computing in Support of Human Resource Management. *Research in Computing Science*, 80, p.107-117, 2014.
- RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: Foundations, D. E. Rumelhart and J. L. McClelland, Eds. Mit Press Computational Models Of Cognition And Perception Series. MIT Press, Cambridge, MA, p. 318-362, 1986.
- RUSSO, S. D.; MIRAGLIA, M.; BORGOGNI, L.; JOHNS, G. How time and perceptions of social context shape employee absenteeism trajectories. *Journal of vocational behavior*, 83(2), 209-217, 2013.

- SADATH, L. Data Mining: A Tool for Knowledge Management in Human Resource. *International Journal of Innovative Technology and Exploring Engineering*, v.2, n.6, p.154-159, 2013.
- SAIDANE, O.; MAHMOUD, I.; GAFSI, L.; HOUDA, A.; TEKAYA, R.; ABDELMOULA, L. Factors leading to work absenteeism in Tunisian ankylosing spondylitis patients. *The Egyptian Rheumatologist*, v.40, n.3, p.183-185, 2018.
- SANTOS, E. J.; QUELUZ, D. P. Factors involved in dentistry absenteeism since the foundation of the Labor Court from 1986 to 2008. *Brazilian Journal of Oral Sciences*, v. 11, n. 4, p. 492-504, 2012.
- SANTOS, T. P. R. Incentivos à doação de sangue: fomento à dignidade da pessoa humana x comercialização inconstitucional. *Direito*, v. 1, n. 12, 2015.
- SASSI, R. J.; SILVA, L. A.; HERNANDEZ, E. del M. A Methodology using Neural Networks to Cluster Validity Discovered from a Marketing Database. In 10th Brazilian Symposium on Artificial Neural Networks (SBRN 2008), Salvador, 27-31 de Outubro. IEEE Proceedings of SBRN 2008. IEEE Computer Society. v.1, p.03-08, 2008.
- SASSI, R. J. Teoria dos rough sets com mapas auto-organizáveis de Kohonen na descoberta de conhecimento em bases de dados. *Revista Produção e Engenharia*, v.2, n.2, p.195-209, 2009.
- SASSI, R, J. Uma arquitetura Híbrida para Descoberta de conhecimento em Bases de Dados: Teoria dos Rough Sets e Redes Neurais Artificiais Mapas Auto-Organizáveis. São Paulo 2006. (Tese de Doutorado em Engenharia Elétrica) – Escola Politécnica da Universidade de São Paulo, 2006.
- SCHAUMBERG, R. L.; FLYNN, F. J. Clarifying the link between job satisfaction and absenteeism: The role of guilt proneness. *Journal of Applied Psychology*, v.102, n.6, p.982–992, 2017.
- SCHULTZ, A. B.; EDINGTON, D. W. Employee health and presenteeism: a systematic review. *Journal of occupational rehabilitation*, v. 17, n. 3, p. 547-579, 2007.
- SHIRAVI, H.; SHIRAVI, A.; GHORBANI, A. A. A survey of visualization systems for network security. *IEEE Transactions on visualization and computer graphics*, v.18, n.8, p.1313-1329, 2012.
- SILVA, D. M. P. P.; MARZIALE, M. H. P. Problemas de Saúde Responsáveis pelo Absenteísmo de Trabalhadores de Enfermagem em um Hospital Universitário. *Acta Scientiarum. Health Sciences*, 25(2), p.191-197, 2012.
- SILVA, I. N.; SPATTI, D.H.; FLAUZINO, R. A. *Redes Neurais Artificiais para Engenharia e Ciências Aplicadas*. SP: Artliber, 2010.
- SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. *Introdução à Mineração de dados: com aplicações em R*. Elsevier. 1ª. ed. Rio de Janeiro, 2016.
- SILVA NETO, M. A.; VILLWOCK, R. S.; SCHEER, S.; STEINER, M. T. A.; DYMINSKI, A. S. Técnicas de mineração visual de dados aplicadas aos dados de instrumentação da barragem de Itaipu. *Visual data mining techniques applied for the analysis of data collected at Itaipu power plant. Gest. Prod.*, São Carlos, v. 17, n. 4, p. 721-734, 2010.

- SILVA, M. M.; CAVEIÃO, C. Análise dos afastamentos de saúde dos trabalhadores de ensino de Divinópolis–MG. *Revista Saúde e Desenvolvimento*, v.10, n.5, p.138-156, 2016.
- SIMÕES, M. G.; SHAW, I. S. *Controle e Modelagem fuzzy*. São Paulo: Blucher: FAPESP, 2007.
- STAVROU, E. T.; CHARALAMBOUS, C.; SPILLOTIS, S. Human resource management and performance: A neural network analysis. *European Journal of Operational Research*, 181, p. 453-467, 2007.
- STROHMEIER, S.; PIAZZA, F. Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications*, 40, p.2410-2420, 2013.
- SUZUKI, T.; MIYAKI, K.; SONG, Y.; TSUTSUMI, A., KAWAKAMI, N.; SHIMAZU, A.; KURIOKA, S. Relationship between sickness presenteeism (WHO–HPQ) with depression and sickness absence due to mental disease in a cohort of Japanese workers. *Journal of affective disorders*, 180, 14-20, 2015.
- TAIFOR, N. A.; ABDULLAH, H. S.; HASHIM, R. Incidence and antecedents of presenteeism: The case of a federal government ministry in Malaysia. In: *Humanities, Science and Engineering (CHUSER), 2011 IEEE Colloquium on. IEEE*, p. 654-659, 2011.
- TAMMITA, U.; SEEDEVI, B. M. D.; JAYARATHNE, D. G. T. S.; WELIANGA, A.; MADUSHANKA, L. H. L. A Study of Employee Absenteeism in the Apparel Industry. *Journal of Human Resource Management*, vol.1, n.1, 2010.
- TAUHID, S., CHU, R., SASANE, R., GLANZ, B. I., NEEMA, M., MILLER, J. R.; KIM, G.; JAMES E. SIGNOROVITCH, B. C. H.; CHITNIS, T.; WEINER, H. L.; BAKSHI, R. Brain MRI lesions and atrophy are associated with employment status in patients with multiple sclerosis. *Journal of neurology*, 262(11), 2425-2432, 2015.
- TAYLOR, P. J. Fifty years ago: Sickness absence resistance. *Occupational Medicine*, v. 68, n. 6, p. 404-404, 2018.
- THANUJA, V.; VENKATESWARLU, B.; ANJANEYULU, G. S. G. N. Applications of data mining in customer relationship management. *Journal of Computer and Mathematical Sciences*, v.2, n.3, p.399-580, 2011.
- THURSTON, E.; GLENDON, A. I. Association of risk exposure, organizational identification, and empowerment, with safety participation, intention to quit, and absenteeism. *Safety Science*, v. 105, p.212-221, 2018.
- TOLBERT, D. V.; McCOLLISTER, K. E.; LeBLANC, W. G.; LEE, D. J.; FLEMING, L. E., MUENNIG, P. The economic burden of disease by industry: Differences in quality-adjusted life years and associated costs. *American journal of industrial medicine*, v.57, n.7, p.757-763, 2014.
- TSUCHIYA, M.; KAWAKAMI, N.; ONO, Y.; NAKANE, Y.; NAKAMURA, Y.; AKIRA FUKAO, A.; TACHIMORI, H.; IWATA, N.; UDA, H.; NAKANE, H.; WATANABE, M.; OORUI, M.; NAGANUMA, Y.; FURUKAWA, T. A.; KOBAYASHI, M.; AHIKO, T.; TAKESHIMA, T.; KIKKAWA, T. Impact of mental disorders on work performance in a community sample of workers in Japan: the World Mental Health Japan Survey 2002–2005. *Psychiatry research*, v.198, n.1, p.140-145, 2012.

- TURBAN, E.; LEIDNER, D.; MCLEAN, E.; WETHERBE, J. (2010). *Tecnologia da Informação para Gestão: Transformando os Negócios na Economia Digital*. (6 ed.). Porto Alegre: Bookman, 2010.
- UCHÔA, J. Q. *Representação e Indução de Conhecimento usando Teoria de Conjuntos Aproximados*. Dissertação de Mestrado. Universidade Federal de São Carlos, 1998.
- URIBE, J. M.; PINTO, D. M.; VECINO-ORTIZ, A. I.; GÓMEZ-RESTREPO, C.; RONDÓN, M. Presenteeism, Absenteeism, and Lost Work Productivity among Depressive Patients from Five Cities of Colombia. *Value in health regional issues*, v.14, p.15-19, 2017.
- VALLE SALVINO, R. *Entre pontes e muros: Tentativas de implantação do correio marítimo entre Portugal e o Brasil na primeira metade do século XVIII*. *Varia História*, v.34, n.64, 2018.
- VERA-CALZARETTA, A.; CARRASCO-DÁJER, C.; DA COSTA, S.; PÁEZ-ROVIRA, D. Factores psicosociales del presentismo en trabajadores del Sistema de Salud chileno. *Revista de Psicología del Trabajo y de las Organizaciones*, 31(2), 119-128, 2015.
- VERSTAPPEN, S. M. M. Rheumatoid arthritis and work: the impact of rheumatoid arthritis on absenteeism and presenteeism. *Best Practice & Research Clinical Rheumatology*, v. 29, n. 3, p. 495-511, 2015.
- VESANTO, J.; ALHONIEMI, E. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, v.11, n.2, p.586-600, 2000.
- VIDAL-GARCÍA, J.; VIDAL, M.; BARROS, R. H. Computational business intelligence, big data, and their role in business decisions in the age of the internet of things. In *Web Services: Concepts, Methodologies, Tools, and Applications*. Chapter 55, p.1048-1067. IGI Global, 2019.
- VISCOVERY. *Viscovery® SOMine*. Disponível em: <<https://www.viscovery.net/>>. Acesso em: 19 jan. 2019.
- XIE, J. L.; JOHNS, G. Interactive effects of absence culture salience and group cohesiveness: a multi-level and cross-level analysis of work absenteeism in the Chinese context, *Journal of Occupational & Organizational Psychology*, v.73, p. 31-52, 2000.
- WADA, K.; ARAKIDA, M.; WATANABE, R.; NEGISHI, M.; SATO, J.; TSUTSUMI, A. The economic impact of loss of performance due to absenteeism and presenteeism caused by depressive symptoms and comorbid health conditions among Japanese workers. *Industrial health*, v.51, n.5, p.482-489, 2013.
- WAN, H. C.; DOWNEY, L. A.; STOUGH, C. Understanding non-work presenteeism: Relationships between emotional intelligence, boredom, procrastination and job stress. *Personality and Individual Differences*, v. 65, p. 86-90, 2014.
- WANG, X.; JIANG, Y. The Study of Enterprise Human Resource Evaluation Research Based on Neural Network. In *Third International Symposium on Information Processing*, p.19- 21, 2010.
- WHO. World Health Organization. Disponível em: <<http://www.who.int/classifications/icd/en/>>. Acesso em: 23 mai. 2019.
- WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

- WU, Y.; WEI, F.; LIU, S.; AU, N.; CUI, W.; ZHOU, H.; QU, H. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, v.16, n.6, p.1109-1118, 2010.
- YAMAMOTO, S.; LOERBROKS, A.; TERRIS, D. D. Measuring the effect of workplace health promotion interventions on “presenteeism”: A potential role for biomarkers. *Preventive medicine*, v. 48, n. 5, p.471-472, 2009.
- YANG, T.; GUO, Y.; MA, M.; LI, Y.; TIAN, H.; DENG, J. Job stress and presenteeism among Chinese healthcare workers: The mediating effects of affective commitment. *International journal of environmental research and public health*, v.14, n.9, p.1-14, 2017.
- YANNIBELLI, V.; AMANDI, A. Maximizing the Effectiveness of Human Resources in the Context of Project Scheduling by an Evolutionary Algorithm. In *First International Conference on Informatics and Computational Intelligence*, p.205-211, 2011.
- YIN, R. K. *Estudo de Caso - Planejamento e Métodos*. Bookman Editora. (5 ed.). Porto Alegre, 2015.
- YU, S.; XIA, J.; YAN, Z.; ZHANG, A., XIA, Y.; GUAN, D.; HAN, J.; WAN, J.; LIU, Y. Loss of work productivity in a warming world: Differences between developed and developing countries. *Journal of Cleaner Production*, v.208, p.1219-1225, 2019.
- YUSUF, O, B.; MUTASIM M. G.; ALI M. S.; KHAN, G.; MAZRAANI, N.; DAWOOD. R.; ABOUZEID, M.; ABDELAAL, M. Pattern and Correlates of Sickness Absenteeism among Workers at the King Abdul Aziz Medical City, Jeddah, Saudi Arabia. *International Journal of Medicine and Medical Sciences*, ISSN: 2051-5731, v.48, n.1, 2015.
- ZHANG, W.; SUN, H.; WOODCOCK, S.; ANIS, A. H. Valuing productivity loss due to absenteeism: firm-level evidence from a Canadian linked employer-employee survey. *Health economics review*, v. 7, n. 1, p.1-14, 2017.
- ZHONG, Y.; MA, A.; SOON ONG, Y.; ZHU, Z.; ZHANG, L. Computational intelligence in optical remote sensing image processing. *Applied Soft Computing*, v.64, p.75-93, 2018.
- ZURADA, J.; KARWOWSKI, W. Knowledge discovery through experiential learning from business and other contemporary data sources: A review and reappraisal. *Information Systems Management*, 28(3), p.258-274, 2011.
- ZURADA, J.; MARKS, R. J.; ROBINSON, J. Review of computational intelligence: imitating life. *Proceedings of the IEEE*. v.83, n.11. (November), p.1588-1592, 1995.
- ZURADA, J. *Introduction to Artificial Neural Systems*; PWS Publishing, 1992.

Anexo A – Relação dos 100 artigos selecionados na revisão bibliométrica

Observou-se que o autor que se destaca na revisão bibliométrica é o professor Gary Johns da Universidade de Concordia no Canadá com 11 artigos, sendo o primeiro artigo publicado em 1978 e o mais recente em 2018, ou seja, 40 anos de pesquisas relacionadas ao estudo do absenteísmo e do presenteísmo. Os seus trabalhos foram destacados em negrito na Tabela do Anexo A. Vale destacar que o professor Gary Johns foi colaborador na realização deste trabalho ao ceder artigos de sua autoria e auxiliando na validação dos atributos para compor a base de dados.

Anexo A – Relação dos 100 artigos selecionados na revisão bibliométrica

Ano	Título do Artigo	Periódico	Referência
2019	Presenteeism: A review and research directions	Human Resource Management Review	Lohaus e Habermann (2019)
2019	Loss of work productivity in a warming world: Differences between developed and developing countries	Journal of Cleaner Production	Yu et al. (2019)
2018	Spillovers from gatekeeping – Peer effects in absenteeism	Journal of Public Economics	Godøy e Dale-Olsen (2018)
2018	Activation against absenteeism –Evidence from a sickness insurance reform in Norway	Journal of Health Economics	Hernæs (2018)
2018	Factors leading to work absenteeism in Tunisian ankylosing spondylitis patients	The Egyptian Rheumatologist	Saidane et al. (2018)
2018	Time Pressure, Time Autonomy, and Sickness Absenteeism in Hospital Employees: A Longitudinal Study on Organizational Absenteeism Records	Safety and Health at Work	Kottwitz et al. (2018)
2018	Predictors of disability and absenteeism in workers with non-specific low back pain: A longitudinal 15-month study	Applied Ergonomics	Lardon et al. (2018)
2018	Association of risk exposure, organizational identification, and empowerment, with safety participation, intention to quit, and absenteeism	Safety Science	Thurston e Glendon (2018)
2018	La gestión de la incapacidad laboral algo más que una cuestión económica	Medicina y Seguridad del Trabajo	Pardo (2018)
2018	Absenteísmo por distúrbios musculoesqueléticos em trabalhadores do Brasil: milhares de dias de trabalho perdidos	Revista Brasileira de Epidemiologia	Haefner et al. (2018)
2018	Mental and behavioral disorders in workers: a study on work leave	Revista da Escola de Enfermagem da USP	Fernandes et al. (2018)
2018	A Job Stress Predictive Model Evaluation Through Classifier's Algorithms	IEEE Latin America Transactions	Rebollar et al. (2018)
2018	A Multi-Level Integration of the Antecedents and Dynamics of Absenteeism	Academy of Management Proceedings	Miraglia e Johns (2018)

2017	Presenteeism, Absenteeism, and Lost Work Productivity among Depressive Patients from Five Cities of Colombia	Journal Value in Health Regional Issues	Uribe et al. (2017)
2017	Drivers of Effort: Evidence from Employee Absenteeism	Journal of Financial Economics	Bennedsen, Tsoutsoura e Wolfenzon (2017)
2017	The pros and cons of sick pay schemes: Testing for contagious presenteeism and noncontagious absenteeism behavior	Journal of Public Economics	Pichler e Ziebarth. (2017)
2017	Sickness absenteeism is associated with sleep problems independent of sleep disorders: results of the 2016 Sleep Health Foundation national survey	Journal of the National Sleep Foundation	Reynolds et al. (2017)
2017	Clarifying the link between job satisfaction and absenteeism: The role of guilt proneness	Journal of Applied Psychology	Schaumberg e Flynn. (2017)
2017	Job stress and presenteeism among Chinese healthcare workers: The mediating effects of affective commitment	International journal of environmental research and public health	Yang et al. (2017)
2017	The conundrum of absenteeism in the Canadian public service: A wicked problem perspective	Canadian Public Administration	McGrandle e Ohemeng (2017)
2017	Absenteísmo na Indústria está associado com o Trabalho em Turnos e com Problemas no Sono	Ciencia & Trabajo	Andrade et al. (2017)
2017	One Hundred Years of Work Design Research: Looking Back and Looking Forward	Journal of Applied Psychology	Parker, Morgeson e Johns (2017)
2016	Absenteísmo por transtornos mentais em trabalhadores de saúde em um hospital no sul do Brasil	Revista Gaúcha de Enfermagem	De Lima Santana et al. (2016)
2016	Why and when do employees imitate the absenteeism of co-workers?	Organizational Behavior and Human Decision Processes	Lieke et al. (2016)
2016	Going to Work III: A Meta-Analysis of the Correlates of Presenteeism and a Dual-Path Model	Journal of Occupational Health Psychology	Miraglia e Johns (2016)
2015	Relationship between sickness presenteeism (WHO-HPQ) with depression and sickness absence due to mental disease in a cohort of Japanese workers	Journal of Affective Disorders	Suzuki et al. (2015)
2015	Sickness presenteeism determines job satisfaction via affective-motivational states	Social Science & Medicine	Karanika-Murray et al. (2015)
2015	Rheumatoid arthritis and work: The impact of rheumatoid arthritis on absenteeism and presenteeism	Best Practice & Research Clinical Rheumatology	Verstappen. (2015)
2015	Factores psicosociales del presentismo en trabajadores del Sistema de Salud chileno	Journal of Work and Organizational Psychology	Vera-Calzaretta et al. (2015)
2015	Brain MRI lesions and atrophy are associated with employment status in patients with multiple sclerosis	J Neurol	Tauhid et al. (2015)
2015	Effect of absenteeism on corporate performance: a case study of cadbury Nigeria plc, Ikeja, Lagos State, Nigeria.	British Journal of Marketing Studies	Onikoyi, Awolusi e Ayodeji. (2015)

2015	Pattern and Correlates of Sickness Absenteeism among Workers at the King Abdul Aziz Medical City, Jeddah, Saudi Arabia	International Journal of Medicine and Medical Sciences	Yusuf et al. (2015)
2014	Understanding non-work presenteeism: Relationships between emotional intelligence, boredom, procrastination and job stress	Personality and Individual Differences	Wan, Downey e Stough (2014)
2014	A dialectical theory of the decision to go to work: Bringing together absenteeism and presenteeism	Human Resource Management Review	Halbesleben, Whitman e Crawford (2014)
2014	Workplace Productivity and Voice Disorders: A Cognitive Interviewing Study on Presenteeism in Individuals With Spasmodic Dysphonia	Journal of Voice	Isetti, e Meyer (2014)
2014	Do the Washington Panel recommendations hold for Europe: investigating the relation between quality of life versus work-status, absenteeism and presenteeism	Cost Effectiveness and Resource Allocation	Knies, Boonen e Severens (2014)
2014	Prediction of Sickness Absenteeism, Disability Pension and Sickness Presenteeism Among Employees with Back Pain	J Occup Rehabil	Bergström et al. (2014)
2014	Development and Measurement Properties of the Dutch Version of the Stanford Presenteeism Scale (SPS-6)	J Occup Rehabil	Hutting et al. (2014)
2014	Influence of Nurse Aide Absenteeism on Nursing Home Quality	The Gerontologist	Castle e Ferguson-Rome (2014)
2014	Presenteeism among self-employed workers: Korean working conditions survey	Annals of Occupational and Environmental Medicine	Kim et al. (2014)
2014	Investigating the Causes of Presenteeism Among Secondary School Employees in Nairobi County	International Journal of Social Sciences Management and Entrepreneurship	Gachui e Were (2014)
2013	The legitimacy of absenteeism from work: a nine nation exploratory study	Cross Cultural Management: An International Journal	Addae, Johns e Boies (2013)
2013	A cross-cultural examination of presenteeism and supervisory support	Career Development International	Lu et al. (2013)
2013	Investigating Two Predictors of Sickness Presenteeism on a Romanian Sample. The Case of Performance-Based Self-Esteem and Overcommitment	Procedia - Social and Behavioral Sciences	Cicei, Mohorea e Teodoru (2013)
2013	Risk Analysis of Hurricane Disruptions on Workforce and Interdependent Regional Sectors	Proceedings of the 2013 IEEE Systems and Information Engineering Design Symposium	Akhtar e Santos, (2013)
2013	Exploration of the effect of workers' influence network on their absence behavior using agent-based modeling and simulation	Proceedings of the 2013 Winter Simulation Conference	Ahn, Anderson e Lee (2013)
2013	Lower Bounds for Estimating Workforce Size in a 24/7 Company	Proceedings of the 2013 IEEE IEEM	Lozano et al. (2013)
2013	Exploratory Study of the Impacts of New Technology Implementation on Burnout and Presenteeism	International Journal of Business and Management	Knani (2013)
2013	Self-assessed mental health problems and work capacity as determinants of return to work: a prospective general population-based study of individuals with all-cause sickness absence	BMC Psychiatry	Hensing (2013)

2013	The Economic Impact of Loss of Performance Due to Absenteeism and Presenteeism Caused by Depressive Symptoms and Comorbid Health Conditions among Japanese Workers	Industrial Health	Wada et al. (2013)
2013	How time and perceptions of social context shape employee absenteeism trajectories	Journal of Vocational Behavior	Russo et al. (2013)
2013	La aparición y declinación del ausentismo como problema para el gobierno de la población trabajadora (Argentina, siglo XX)	Trabajo y Sociedad	Haidar (2013)
2012	Transformational and passive avoidant leadership as determinants of absenteeism	Leadership & Organization Development Journal	Frooman, Mendelson e Kevin (2012)
2012	Factors involved in dentistry absenteeism since the foundation of the Labor Court from 1986 to 2008	Brazilian Journal of Oral Sciences	Santos e Queluz (2012)
2012	The case for presenteeism — Evidence from Norway's sickness insurance program	Journal of Public Economics	Markussen, Mykletun e Røed (2012)
2012	Impact of mental disorders on work performance in a community sample of workers in Japan: The World Mental Health Japan Survey 2002–2005	Psychiatry Research	Tsuchiya et al., (2012)
2012	The Use and Performance of Productivity Scales to Evaluate Presenteeism in Mood Disorders	Value in Health	Despiégel et al. (2012)
2012	Application of a neuro fuzzy network in prediction of absenteeism at work	Conference Information Systems and Technologies	Martiniano et al. (2012)
2012	Factors associated with absenteeism, presenteeism and activity impairment in patients in the first years of RA	Rheumatology	Bansback et al. (2012)
2012	Aversive Workplace Conditions and Absenteeism: Taking Referent Group Norms and Supervisor Support Into Account	Journal of Applied Psychology	Biron e Bamberger (2012)
2011	Mental disorders as a cause of absenteeism among public workers in São Paulo	Archives of Clinical Psychiatry	Cruz et al. (2011)
2011	O impacto da disфонia em professores: queixas vocais, procura por tratamento, comportamento, conhecimento sobre cuidados com a voz, e absenteísmo	Rev Soc Bras Fonoaudiol	Mestre e Ferreira (2011)
2011	Problemática multifactorial del absentismo laboral, el presentismo y la procrastinación en las estructuras en que se desenvuelve el trabajador	Medicina y Seguridad del Trabajo	García (2011)
2011	Factors associated with presenteeism among employed Australian adults reporting lifetime major depression with 12-month symptoms	Journal of Affective Disorders	Cocker et al. (2011)
2011	Development of Model of Workers Mental Processes Related to Absence Norm as Behavior Rule in Agent-Based Simulation	Proceedings of the 2011 Winter Simulation Conference	Ahn e Lee (2011)
2011	Incidence and Antecedents of Presenteeism: The case of a Federal Government Ministry in Malaysia	Colloquium on Humanities, Science and Engineering Research	Taifor, Abdullah e Hashim (2011)
2011	Predicting failure to return to work	Internal Medicine Journal	Mills (2011)
2011	Does Physical Activity Impact on Presenteeism and Other Indicators of Workplace Well-Being?	Sports Med	Brown et al. (2011)
2011	Absenteeism in a Represented Environment	International Journal of Humanities and Social Science	Romero e Strom (2011)

2011	Relative magnitude of presenteeism and absenteeism and work-related factors affecting them among health care professionals	Int Arch Occup Environ Health	Rantanen e Tuominen (2011)
2011	Attendance Dynamics at Work: The Antecedents and Correlates of Presenteeism, Absenteeism, and Productivity Loss	Journal of Occupational Health Psychology	Johns (2011)
2010	A influência da satisfação profissional no absentismo laboral	Rev Port Saúde Pública	Agapito e Sousa (2010)
2010	An optimization model for task assignment in home health care	Health Care Management (WHCM), 2010 IEEE Workshop on	Bachouch, Guinet e Hajri-Gabouj (2010)
2010	Presenteeism: Are We Hurting the Patients We are Trying to Help?	J Gen Intern Med	Landry e Miller, (2010)
2010	Depression-related absenteeism and presenteeism amongst employed Australian adults: Do working conditions influence the health and economic consequences?	Journal of Affective Disorders	Cocker et al. (2010)
2010	Predictors of Sickness Absence and Presenteeism: Does the Pattern Differ by a Respondent's Health?	Journal Occupational and Environmental Medicine	Böckerman e Laukkanen (2010)
2010	Presenteeism in the workplace: A review and research agenda	Journal of Organizational Behavior	Johns (2010)
2010	A Study of Employee Absenteeism in the Apparel Industry	Human Resource Management Journal	Tammita et al. (2010)
2009	Welfare wars: public service frontline absenteeism as collective resistance	Qualitative Research in Accounting & Management	Junor, O'brien e O'donnell (2009)
2009	Measuring the effect of workplace health promotion interventions on "presenteeism": A potential role for biomarkers	Preventive Medicine	Yamamoto, Loerbroks e Terris, (2009)
2009	Impact of gastroesophageal reflux disease on work absenteeism, presenteeism and productivity in daily life: a European observational study	Health and Quality of Life Outcomes	Gisbert et al. (2009)
2009	Presenteeism in Finland: Determinants by Gender and the Sector of Economy	Ege Academic Review	Böckerman e Laukkanen (2009)
2008	Factors in absenteeism and presenteeism: life events and health events	Management Research News	Macgregor, Cunningham e Caverley (2008)
2008	Working conditions that contribute to absenteeism among nurses in a provincial hospital in the Limpopo Province	Curationis	Nyathi e Jooste (2008)
2008	Inactividad Física y Ausentismo en el Ámbito Laboral	Rev. salud pública	Martínez-López e Saldarriaga-Franco (2008)

2008	Pandemic Influenza, Worker Absenteeism and Impacts on Freight Transportation	Proceedings of the 41st Hawaii International Conference on System Sciences	Jones et al. (2008)
2008	Present but sick: a three-wave study on job demands, presenteeism and burnout	Career Development International	Demerouti et al. (2008)
2008	Work Strain, Health, and Absenteeism: A Meta-Analysis	Journal of Occupational Health Psychology	Darr e Johns (2008)
2008	Organizational Wellness Programs: A Meta-Analysis	Journal of Occupational Health Psychology	Parks e Steelman (2008)
2007	Employee Health and Presenteeism: A Systematic Review	J Occup Rehabil	Schultz e Edington (2007)
2007	Predicting absenteeism and turnover intentions by past absenteeism and work attitudes	Career Development International	Cohen e Golan (2007)
2006	Prevalencia, impacto en la productividad y costos totales de las principales enfermedades en los trabajadores de un hospital al sur del Perú en el año 2003.	Rev Med Hered	Gutiérrez et al. (2006)
2005	Choosing to work when sick :workplace presenteeism	Social Science & Medicine	Dew, Keefe e Small (2005)
2004	30 años de estudio sobre ausentismo laboral en Chile: una perspectiva por tipos de empresas	Rev Méd Chile	Mesa e Kempffer (2004)
2003	How methodological diversity has improved our understanding of absenteeism from work	Human Resource Management Review	Johns (2003)
1991	Substantive and Methodological Constraints on Behavior and Attitudes in Organizational Research	Organizational Behavior and Human Decision Processes	Johns (1991)
1978	Attitudinal and Nonattitudinal Predictors of Two Forms of Absence from Work	Organizational Behavior and Human Performance	Johns (1978)
1975	The Management of Absenteeism	The Production Engineer	Almond (1975)
1968	Labour—Turnover and Absenteeism	The Production Engineer	Chan (1968)
1945	Co-Operation Between the Electrical and Medical Professions, and Possible Effect on Absenteeism In Industrial Undertakings	Electrical Engineering	Miller (1945)

Anexo B – Certificado de Registro de Programa de Computador



REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA INDÚSTRIA, COMÉRCIO EXTERIOR E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512018001297-3**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 14/01/2018, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Software Tool for Absenteeism Prediction

Data de criação: 14/01/2018

Titular(es): ASSOCIAÇÃO EDUCACIONAL NOVE DE JULHO

Autor(es): RENATO JOSÉ SASSI; RICARDO PINTO FERREIRA; ANDRÉA MARTINIANO DA SILVA

Linguagem: C

Campo de aplicação: AD-02

Tipo de programa: IA-02

Algoritmo hash: SHA-512

Resumo digital hash:
0ad5657fafa7b1cdd5b51239c7a345d824d7e408b351d41833dfda2efd691560f40d550aec07781c47419565da1ecfba64d
a58dfbfa507c9f86c165f067e0c5a

Expedido em: 31/07/2018

15 de Novembro
REPÚBLICA FEDERATIVA DO BRASIL
de 1889