UNIVERSIDADE FEDERAL DO MARANHÃO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LAUDELINO SIMÃO CAMPOS ALMEIDA FILHO

ANÁLISE DE OCORRÊNCIAS DE DADOS DE TRÂNSITO EM VIAS TERRESTRES

SÃO LUÍS

2018

LAUDELINO SIMÃO CAMPOS ALMEIDA FILHO

ANÁLISE DE OCORRÊNCIAS DE DADOS DE TRÂNSITO EM VIAS TERRESTRES

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Ivo José da Cunha Serra

SÃO LUÍS

2018

Laudelino Simão Campos Almeida Filho.

Análise De Ocorrências De Dados De Trânsito Em Vias Terrestres/Laudelino Simão Campos Almeida Filho – São Luís, 2018.
34 p.

Orientador. Prof. Dr. Ivo José da Cunha Serra

Monografia (Graduação) – Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnológicas
Curso de Graduação em Ciência da Computação, 2018

1 Introdução. 2 Fundamentação Teórica. 3 Análise De Ocorrências De Dados De Trânsito Em Vias Terrestres. I. Ivo José da Cunha Serra. II. Universidade Federal do Maranhão. III. Ciência da Computação. IV. Título

LAUDELINO SIMÃO CAMPOS ALMEIDA FILHO

ANÁLISE DE OCORRÊNCIAS DE DADOS DE TRÂNSITO EM VIAS TERRESTRES

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Aprovada em 09/11/2018

BANCA EXAMINADORA

Prof. Dr. Ivo José da Cunha Serra (Orientador)
Centro de Ciências Exatas e Tecnologia – CCET
Universidade Federal do Maranhão – UFMA

Prof. MSc. Carlos Eduardo Portela Serra de Castro
Centro de Ciências Exatas e Tecnologia – CCET
Universidade Federal do Maranhão – UFMA

Prof. Dr. Tiago Bonini Borchartt
Centro de Ciências Exatas e Tecnologia – CCET
Universidade Federal do Maranhão - UFMA

“Portanto dEle, por Ele e para Ele são todas as coisas. A Ele seja a glória perpetuamente! “

Rm 11:36.

AGRADECIMENTOS

Em primeiro lugar à Deus, por ser meu baluarte, refúgio em tempo todo, que me amou primeiro enviando seu Filho amado para a remissão dos meus pecados.

Aos meus pais, Laudelino Almeida e Hermínia Ribeiro Medeiros, que sempre confiaram em mim, me encorajando com palavras e ações de perpétua valentia. À Samara, minha companheira enviada de Deus para mim, amor da minha vida, que sempre esteve comigo em todos os momentos: da mais profunda agonia da alma até o mais momento de júbilo e alegria.

Aos meus amigos que sempre me ajudaram a crescer, tanto moralmente quanto academicamente, em especial o Samir, que sempre demonstrou ser um amigo fiel em todas as horas.

Ao meu orientador, professor Ivo José da Cunha Serra, pela paciência e disponibilidade neste trabalho.

RESUMO

Devido ao crescimento contínuo de dados computacionais que são produzidos e armazenados, técnicas de mineração de dados tornaram-se cada vez mais necessárias para a procura de padrões relevantes de informações nestes grandes volumes. A mineração de dados surge com o propósito de identificar e extrair informações relevantes de bases de dados. Este trabalho descreve uma análise de ocorrências de dado em vias terrestres, elucidando a fundamentação teórica de mineração de dados, classificação de dados e o classificador usado: Multi Layer Perceptron. São discutidos: a descrição dos dados utilizados pela classificação e a definição do *dataset* utilizado. Sugestões de aplicações dos resultados encontrados são feitas, exibindo informações encontradas sobre dados de trânsito, corroborando para possíveis melhorias de trânsito da parte das autoridades competentes.

Palavras-chave: Mineração de Dados, Classificação, MLP(Multi Layer Perceptron).

ABSTRACT

Due to the continued growth of computational data that is produced and stored, data mining techniques have become increasingly necessary to search for relevant patterns of information in these large volumes. The data mining sufficiently deal the purpose of identifying and extracting relevant information based on this database. This monograph describes an analysis of data occurrences on land routes, elucidating the theoretical foundations of data mining, data classification and classifier used: Multi-Layer Perceptron. The description of the data used by the classification and the definition of the dataset used is discussed. Found applications of the results of suggestions are made in the last section of Chapter 3, displaying information found on traffic data, corroborating for possible transit improvements of the competent authorities.

Keywords: Data Mining, Classification, MLP (Multi-Layer Perceptron).

Lista de Figuras

Lista de Gráficos

Lista De Abreviaturas E Siglas

Arff	Attribute-Relation File Format
CTB	*Código de Trânsito Brasileiro*
CSV	*Comma-separated values*
KDD	*Knowledge Discovery in Databases*
MLP	*Multi Layer Perceptron*
PRF	*Polícia Rodoviária Federal*
RS	*Rio Grande do Sul*

Sumário

1 INTRODUÇÃO

O trânsito é de suma importância para a mobilização e transporte de pessoas e mercadorias, além de possibilitar uma maior efetivação de serviços, de modo a otimizar a locomoção no dia-a-dia das pessoas. Mas, o que é trânsito? Segundo o Código de Trânsito Brasileiro (CTB)[1], no Artigo 1º, diz que trânsito: *"É a utilização das vias por veículos motorizados, veículos não motorizados, pedestres e animais, para fins de circulação, parada ou estacionamento e operação de carga ou descarga."*.

Segundo o Detran do Rio Grande do Sul[2], a cada 15 minutos, um motorista gaúcho tem suspenso o direito de dirigir, acumulando o total de 42,762 habilitados suspensos só do período de 2007 a 2013. A principal causa dessa suspensão é o excesso de velocidade, isso porquê, só em 2007, no total de 3 milhões de infrações, 611.565 (20,40%) infrações foram por excesso de velocidade. Já em 2013, do total de 4 milhões de infrações autuadas, 975.713 (24,40%) foram por excesso de velocidade, tendo o aumento de 364.148 infrações desse tipo.

Com o latente crescimento da tecnologia da informação, novas técnicas foram criadas com o intuito de facilitar o uso da informação de modo mais automatizado, com a finalidade de organizar os dados de forma mais acessível, tanto a usuários leigos quanto à experientes profissionais da área de tecnologia.

Nesse escopo, a Mineração de Dados, surgiu da necessidade de processar grandes volumes de dados, onde tem como objetivo descobrir padrões úteis e recentes, que poderiam de alguma outra forma, permanecer ignorados e não explorados devidamente (PANG-NING et al., 2009).

Com base nesses dados alarmantes apresentados sobre infrações de trânsito, podemos extrair dados de Ocorrência de Infrações em Vias Terrestres utilizando a Mineração de Dados. Por isso, este trabalho tem o objetivo de abordar uma discussão sobre o uso de classificação de dados com o intuito de

[1] **https://www.jusbrasil.com.br/topicos/10632340/artigo-1-da-lei-n-9503-de-23-de-setembro-de-1997**

[2] **http://www.detran.rs.gov.br/conteudo/29998/perigo-nas-estradas%3A-infratores-sao-homens,-entre-26-e-30-anos**

encontrar padrões de infrações de trânsito, partindo do embasamento teórico-científico até a aplicação prática dos mesmos.

Nesse contexto será abordada, a partir da aplicação prática do classificador apresentado *Multi Layer Perceptron* (MLP), uma análise de ocorrências de dados de trânsito (infrações) em vias terrestres na BR 116

1.1 OBJETIVOS

O objetivo geral deste trabalho é prover sugestões de aplicações dos dados encontrados de infrações de trânsito em via terrestre brasileira por meio da aplicação de uma técnica de classificação de dados.

1.1.1 Objetivos específicos

• Minerar informações de infrações de trânsito para a aplicação prática da análise de infrações de trânsito em vias terrestres;

• Elucidar os resultados encontrados da classificação de dados;

• Discutir sobre informações encontradas sobre infrações em vias terrestres a partir de sua classificação, e possíveis aplicações desse resultado.

1.2 ORGANIZAÇÃO DO TRABALHO

Esta monografia apresenta a seguinte organização:

No Capítulo 2, Fundamentação Teórica, seguem informações importantes para o contexto e entendimento do trabalho, tais como o entendimento sobre conceitos, técnicas e métodos de Mineração de Dados e o embasamento teórico do Classificador *Multi Layer Perceptron* (MLP);

No Capítulo 3, Análise De Ocorrências De Dados De Trânsito Em Vias Terrestres, é apresentada a descrição e o uso prático do do dataset utilizado; a descrição sobre o domínio de ocorrências das infrações de trânsito; sugestões de aplicações dos resultados encontrados, e informações relevantes obtidas da classificação para aplicação prática no dia-a-dia.

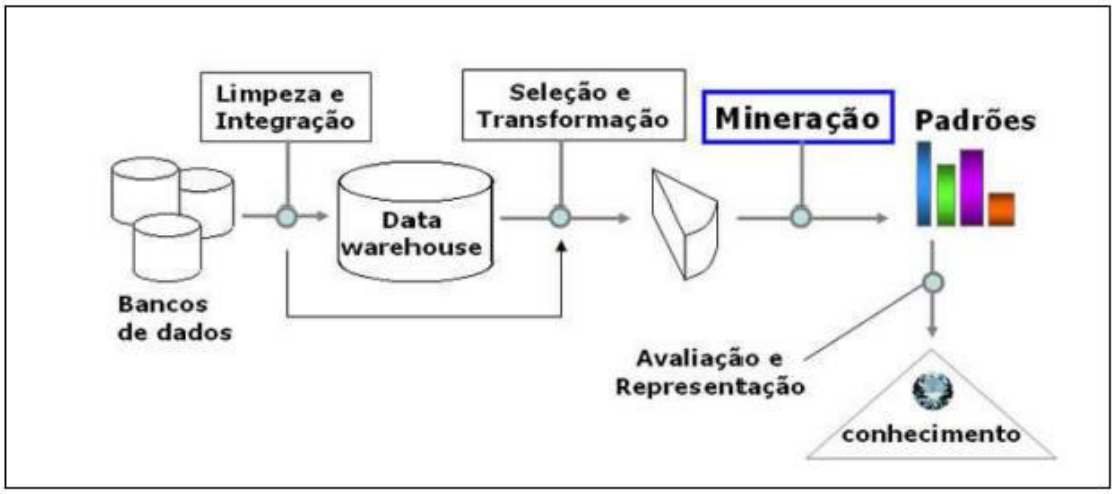
No Capítulo 4, Conclusão e Trabalhos Futuros, apresentam-se as considerações finais e possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Mineração de Dados refere-se a disciplina que tem como objetivo descobrir “novas” informações através da análise de grandes quantidades de dados. A mineração de dados pode ser considerada como uma parte do processo de Descoberta de Conhecimento em Banco de Dados (KDD - Knowledge Discovery in Databases) (BAKER et al., 2010). Entende-se que a mineração de dados tem o papel de filtrar dados, do escopo mais abrangente ao mais específico, com o intuito de qualificar a informação. Devido ao grande avanço qualitativo dessa disciplina, o uso da mineração de dados foi abrangido não somente a áreas científicas, mas também na área financeira, comercial, de marketing, medicina entre outras.

Para que a Mineração de Dados seja feita de forma produtiva, é necessário o cumprimento de etapas, como explicitada na Figura 1

Figura 1 – Etapas do processo



Fonte: (AMO, 2004).

Limpeza e integração: Trata-se da retirada de dados inconsistentes para a melhora da aplicação dos algoritmos.

Seleção: etapa na qual é selecionado os parâmetros que o usuário julga necessário para o seu objetivo principal.

Transformação: esta etapa ocorre após a seleção e limpeza dos dados, trata-se da transformação dos dados em dados compatíveis com os algoritmos de mineração.

Mineração: etapa na qual ocorre a utilização das técnicas de mineração para extrair padrões.

2.1 TAREFAS TRADICIONAIS DE MINERAÇÃO DE DADOS

A mineração de dados vem se expandindo no decorrer do tempo, se desenvolvendo nos mais diversos tipos de ambientes como nos negócios, bolsa de valores entre outros seguimentos. Segundo (HARRISON, 1998) as tarefas foram elaboradas no início do processo de mineração de dados para nortear os algoritmos de acordo com o tipo de conhecimento extraído, ou seja, caso se queira realizar uma predição de um fato futuro com base no passado, a tarefa a ser utilizada seria a classificação, porém não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes tarefas servem para diferentes propósitos, a seguir veremos uma descrição das tarefas mais tradicionais, abordando em alguns casos suas aplicações.

. Em relação aos métodos ou técnicas para mineração de dados, tem-se, como exemplo: árvores de decisão, regras de associação, redes neurais, máquinas de vetores de suporte, algoritmos genéticos entre outros. Há diversos, devido ao grande desenvolvimento deste campo, entretanto, o foco deste estudo de caso é o da classificação usando Rede Neural, *Multi Layer Perceptron*, que será explanada na Seção 2.3

2.1.1 Classificação

Uma das tarefas mais tradicionais, a classificação tem como objetivo identificar a qual classe um determinado registro pertence. Nesta tarefa, primeiramente há a indução do classificador a partir de um conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de "induzir"como classificar um novo registro (aprendizado supervisionado), tendo assim a classificação dos registros. Por exemplo, para categorizar cada registro de um conjunto de dados contendo as informações sobre os colaboradores de três classes: Perfil Técnico, Perfil Negocial e Perfil Gerencial.

A partir daí, é realizado uma análise dos registros, posteriormente é possível estimar em qual categoria um novo colaborador ao ser cadastrado irá

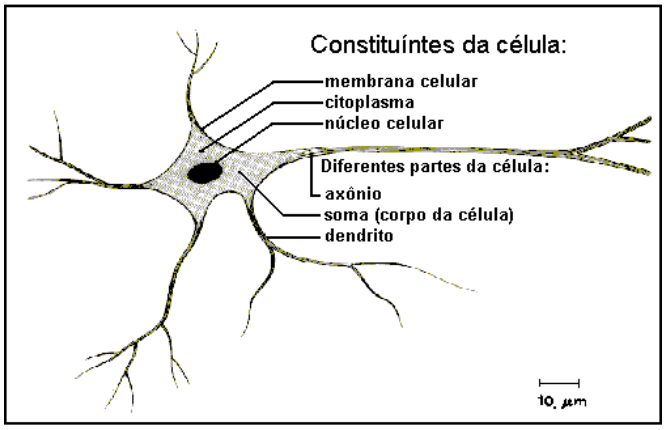
ser alocado, (CAMILO; SILVA, 2009). Na Seção 2.2 alguns métodos de classificação bastante utilizados serão abordados. A tarefa de classificação pode ser usada por exemplo para:

- Determinar quando uma transação de cartão de crédito pode ser uma fraude;
- Detectar quando uma mensagem é spam em e-mails
- Diagnosticar quais doenças podem se manifestar em um paciente baseado em seus exames de rotina;
- Identificar ações de uma determinada pessoa como os sites que costuma visitar ou programas que costuma baixar, baseados nestas informações pode-se detectar se esta pessoa é ou não uma ameaça para a segurança.

2.1.1.1 Redes Neurais

O sistema nervoso é formado por um conjunto extremamente complexo de células, os neurônios. Eles têm um papel essencial na determinação do funcionamento e comportamento do corpo humano e do raciocínio. Os neurônios (Figura 2) são formados pelos dendritos, que são um conjunto de terminais de entrada, pelo corpo central, e pelos axônios que são longos terminais de saída.

Figura 2 – Exemplo de Neurônio Natural



Fonte: (PANG-NING et al., 2009)

Os neurônios se comunicam através de sinapses. Sinapse é a região onde dois neurônios entram em contato e através da qual os impulsos nervosos são

transmitidos entre eles. Os impulsos recebidos por um neurônio A, em um determinado momento, são processados, e atingindo um dado limiar de ação, o neurônio A dispara, produzindo uma substância neurotransmissora que flui do corpo celular para o axônio, que pode estar conectado a um dendrito de um outro neurônio B. O neurotransmissor pode diminuir ou aumentar a polaridade da membrana pós-sináptica, inibindo ou excitando a geração dos pulsos no neurônio B. Este processo depende de vários fatores, como a geometria da sinapse e o tipo de neurotransmissor.

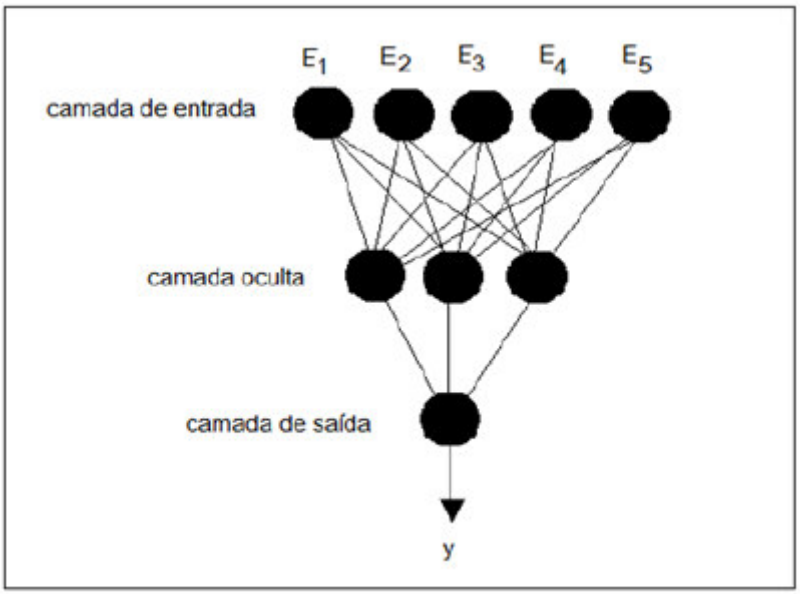
Em média, cada neurônio forma entre mil e dez mil sinapses. O cérebro humano possui cerca de 10 E11 neurônios, e o número de sinapses é de mais de 10 E14, possibilitando a formação de redes muito complexa.

Já as Redes Neurais Artificiais foram originalmente projetadas por psicólogos e neurobiologistas que procuravam desenvolver um conceito de neurônio artificial análogo ao neurônio natural. Intuitivamente, uma rede neural artificial é um conjunto de unidades do tipo entrada e saída, tais unidades são conectadas umas às outras e cada conexão tem um peso associado.

Cada unidade representa um neurônio. Os pesos associados a cada conexão entre os diversos neurônios é um número entre -1 e 1 e mede de certa forma qual a intensidade da conexão entre os dois neurônios (AMO, 2004).

De acordo com PANG-NING et al.(2009). A Rede Neural Artificial é dividida em dois tipos de modelos: *perceptron* que se trata de um modelo mais simples, este modelo apresenta um conjunto de nodos de entrada que são usados para representar os atributos de entrada e um nodo se saída, que tem por finalidade, representar a saída do modelo, o outro modelo é mais complexo e é chamado de rede neural artificial multicamadas (*Multi Layer Perceptron*), esta rede pode conter diversas camadas intermediárias entre suas camadas de entrada e saída, além disso, apresentam um conjunto de unidades de entrada em forma de um vetor de entradas E1,E2,E3,E4,E5 com pesos relacionados a cada entrada, conforme a Figura 3, o sinal de entrada passa pela camada oculta (camadas intermediárias) até chegar ao nodo de saída.

Figura 3 – Exemplo de Rede Neural

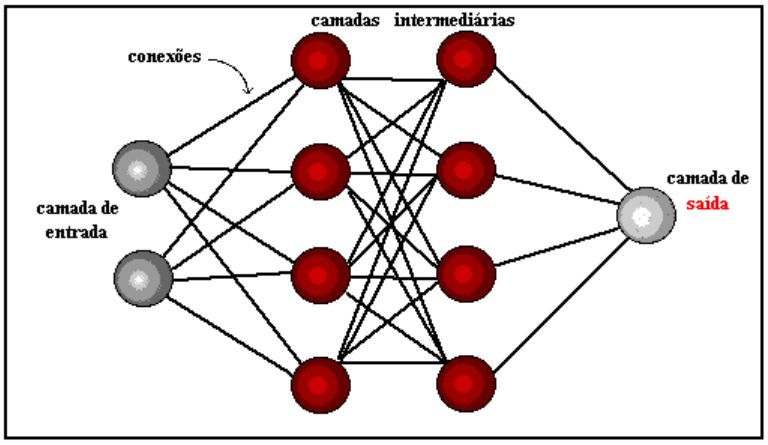


Fonte: (PANG-NING et al., 2009) Adaptado.

A maioria dos modelos de redes neurais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados. Em outras palavras, elas aprendem através de exemplos.

Arquiteturas neurais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior, como mostrado na Figura 4

Figura 4 – Estrutura de um Perceptron Multi Camadas



Fonte: (CAMILO; SILVA, 2009)

Usualmente as camadas são classificadas em três grupos: Camada de Entrada: onde os padrões são apresentados à rede; Camadas Intermediárias ou

Escondidas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características; Camada de Saída: onde o resultado final é concluído e apresentado.

O processo de aprendizado de um certo conceito pela rede neural corresponde à associação de pesos adequados às diferentes conexões entre os neurônios, desta forma, utilizando redes neurais é possível treinar e classificar a partir das entradas, que são os atributos do modelo que queremos obter a classificação. Logo, entende-se que a propriedade mais importante das redes neurais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento.

O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas. Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Outro fator importante é a maneira pela qual uma rede neural se relaciona com o ambiente. Nesse contexto existem os seguintes paradigmas de aprendizado:

- *Aprendizado Supervisionado*: quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada;

- *Aprendizado Não Supervisionado (auto-organização)*: quando não existe um agente externo indicando a resposta desejada para os padrões de entrada;

- *Reforço:* quando um crítico externo avalia a resposta fornecida pela rede.

Denomina-se ciclo uma apresentação de todos os N pares (entrada e saída) do conjunto de treinamento no processo de aprendizado. A correção dos pesos num ciclo pode ser executada de dois modos:

1) Modo Padrão: A correção dos pesos acontece a cada apresentação à rede de um exemplo do conjunto de treinamento. Cada correção de pesos baseia-se somente no erro do exemplo apresentado naquela iteração. Assim, em cada ciclo ocorrem N correções.

2) Modo Batch: Apenas uma correção é feita por ciclo. Todos os exemplos do conjunto de treinamento são apresentados à rede, seu erro médio é calculado e a partir deste erro fazem-se as correções dos pesos.

2.2 CLASSIFICADOR *MULTI LAYER PERCEPTRON OU PERCEPTRON MULTI-CAMADAS*

Quando Redes Neurais Artificiais de uma só camada são utilizadas os padrões de treinamento apresentados à entrada são mapeados diretamente em um conjunto de padrões de saída da rede, ou seja, não é possível a formação de uma representação interna. Neste caso, a codificação proveniente do mundo exterior deve ser suficiente para implementar esse mapeamento.

Tal restrição implica que padrões de entrada similares resultem em padrões de saída similares, o que leva o sistema à incapacidade de aprender importantes mapeamentos. Como resultado, padrões de entrada com estruturas similares, fornecidos do mundo externo, que levem a saídas diferentes não são possíveis de serem mapeados por redes sem representações internas, isto é, sem camadas intermediárias.

Assim temos o uso da Multi Layer Perceptron , que é um classificador que, como seu próprio nome conceitua, utiliza várias camadas de neurônios ocultas internas para alcançar um resultado satisfatório de classificação, pois possui a capacidade de generalização de uma função (ou classe) com alto nível de precisão, pois utiliza o conceito de retro propagação para agrupar os dados da forma mais semelhante possível, além de serem úteis na pesquisa em termos de sua capacidade de resolver problemas estocásticos, que muitas vezes permitem obter soluções aproximadas.

No entanto problemas como o overfitting, ou sobre-ajuste, (onde o classificador se torna “viciado” em classificar baseado somente no mesmo conjunto de dados, tendo dificuldade de classificar novos conjuntos de dados não conhecidos anteriormente) e underfitting ou sub-ajuste (que se comporta de forma contrária ao overfitting, ou seja, o classificador classifica com baixa precisão o conjunto de dados, pois não se especializa no próprio conjunto de

dados) podem comprometer o desempenho e a descoberta de dados de forma produtiva.

Então a solução é treinar o classificador com a técnica de aprendizagem supervisionada (retro propagação). O treinamento supervisionado da rede MLP utilizando retro propagação consiste em dois passos : No primeiro, um padrão (o conjunto de dados de infrações) é apresentado às unidades da camada de entrada e, a partir desta camada as unidades calculam sua resposta que é produzida na camada de saída(a classificação em si), o erro é calculado e no segundo passo, este é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados (treinados até chegarem à classificação).

Porém, temos como consequência que, as redes neurais que utilizam retro propagação, assim como muitos outros tipos de redes neurais artificiais, podem ser vistas como "caixas pretas", na qual quase não se sabe porque a rede chega a um determinado resultado, uma vez que os modelos não apresentam justificativas para suas respostas. Neste sentido, muitas pesquisas vêm sendo realizadas visando a extração de conhecimento de redes neurais artificiais, e na criação de procedimentos explicativos, onde se tenta justificar o comportamento da rede em determinadas situações.

Uma outra limitação refere-se ao tempo de treinamento de redes neurais utilizando retro propagação que tende a ser muito lento. Algumas vezes são necessários milhares de ciclos para se chegar à níveis de erros aceitáveis, principalmente se estiver sendo simulado em computadores seriais, pois a CPU deve calcular as funções para cada unidade e suas conexões separadamente, o que pode ser problemático em redes muito grandes, ou com grande quantidade de dados. Porém, esses problemas não serviram de barreira para utilizar o MLP, pois, devido o interesse resultante das aulas de Mineração de Dados na Faculdade, surgiu a curiosidade em mim de trabalhar com o MLP à cunho experimental para esta Análise de Dados.

3 – ANÁLISE DE OCORRÊNCIAS DE DADOS DE TRÂNSITO EM VIAS TERRESTRES

Informações interessantes e valiosas como, por exemplo, que tipo de carros mais cometem infrações dado um determinado turno, ou qual marca de carro é mais catalogada como infratora, ou se existe um número maior de concentração de casos de alcoolemia relacionados à uma determinada espécie de veículo, são possíveis de serem adquiridas e catalogadas devido à classificação de dados.

Neste capítulo, será abordado um estudo de caso sobre classificações de infrações em vias terrestres, com a finalidade de disponibilizar informações relevantes neste domínio, descobrindo como os dados se comportam, como exemplo: alto índice de excesso de velocidade no turno da manhã.

Das métricas gerais utilizadas para avaliar classificadores temos: Taxa de erro (número de erros / total), Validação Cruzada (consiste em dividir os dados em n partições de treino e teste, e calcular o erro médio) e a Matriz de Confusão (Definida através de Falsos Positivos, Falsos Negativos, Positivos e Negativos).

Nesse estudo de caso, os resultados obtidos da classificação foram avaliados e dispostos nas formas de Taxa de Erro e de Matriz de Confusão, devido à sua fácil e intuitiva interpretação gráfica.

3.1 DESCRIÇÃO DOS DADOS À SEREM EXPLORADOS PARA CLASSIFICAÇÃO

Foi utilizado um dataset disponibilizado no site da Polícia Rodoviária Federal[1] (PRF), no qual todas as informações coletadas na ocasião de uma autuação de trânsito em rodovias federais são catalogadas, resguardando-se os dados pessoais, que permitem a identificação do infrator.

O dataset é formado por 13 semestres (de 2007 a 2013) de abordagens policiais a motoristas infratores (motos, carros, caminhões), onde os tipos de

[1] http://dados.gov.br/dataset/multas-rodovias-federais

autuações podem ser com abordagem (alcoolemia e velocidade acima da permitida) ou sem abordagem (velocidade acima da permitida).

Cada semestre conta com cerca de um milhão de entradas e possui 36 atributos, sendo eles :

cod_status	num_auto	tip_abordagem
tip_cnh_condutor	uf_cnh_condutor	ind_assinou_auto
num_br_infracao	uf_infracao	num_km_infracao
cod_municipio_inf	hor_infracao	ind_sentido_trafego
tip_medicao	med_realizada	lim_regulamentar
med_considerada	exc_verificado	ind_observacao
dat_digitacao	txt_data_infracao	dat_infracao
cod_ultima_inf	ind_laudo_medico	nom_modelo_veiculo
nom_especie_vei	nom_unidade	dat_afericao
dat_env_correio	nom_imagem	nom_imagem_tratada
cod_veiculo_marca	cod_veiculo	nome_veiculo_marca
metragem_infracao	ind_veiculo	

Constam vários dados a respeito das multas aplicadas, dos veículos autuados e das condições do local, entre outros. Entre os atributos mais relevantes, temos: o número do km da infração, o estado e a altura da rodovia onde foi cometida a infração, o tipo e o turno da infração, a marca e o tipo do veículo.

Por ser um dataset tão grande (número muito grande de atributos), foi necessário limitar o número de atributos (para 4) à serem utilizados nos classificadores para um melhor desempenho computacional.

Devido ao dataset abranger dados de todo o território nacional, foi necessário diminuir o escopo do estudo de caso para tornar mais adequado à uma monografia devido à falta de recursos computacionais suficientes para computar todos esses dados, limitando-o a uma rodovia. Ainda para que pudesse manter as informações sobre onde foi cometida determinada infração, foi preciso

definir uma rodovia e um estado por onde essa rodovia passa, já que a cada entrada em um novo estado, a contagem dos quilômetros é reiniciada.

Assim, foi escolhida a BR-116 (começa no Ceará e termina no Rio Grande do Sul, passando por Rio de Janeiro e São Paulo), uma das mais movimentadas e conhecidas do pais. Foi escolhido o estado com maior número de infrações desta rodovia, o Rio Grande do Sul, totalizando 1380 instâncias de infrações. Foram classificadas todas as 1380 Instâncias, baseadas nos seguintes atributos:

- “tip-medicao" (Tipo da Infração – Excesso de velocidade, Alcoolemia, Dimensão e Peso)
- "nome-veiculo-marca" (Nome do automóvel)
- "turno" (Madrugada, Manhã, Tarde, Noite)
- “nome-veiculo-especie" (Carga, Especial, Passageiro, Tração, Misto)

3.2 DEFINIÇÃO E DESCRIÇÃO DO DATASET EM TERMOS DE OBJETOS E ATRIBUTOS

Originalmente os dados estavam em um arquivo no formato comma-separated values (CSV). No ambiente (Weka), no qual foi trabalhado o projeto, este formato não permitia a definição de alguns tipos de dados originais do arquivo.csv (txt_data; ind_observ; hora_infracao) aceitando apenas tipo numérico e nominal. Para isso, foi necessário a conversão do formato original para o formato Attribute-Relation File Format (arff), que é o formato padrão do ambiente. Este formato (.arff) permite definirmos datas e horários reconhecidos no Weka.

Entretanto, a saída para definir a hora da infração, ainda não era a ideal, já que o formato se encontra em milissegundos. Dessa forma foi necessário um pré-processamento, utilizando uma função do MS Excel para transformar o atributo hora_infracao para turno (madrugada, manhã, tarde, noite). Isso foi de bastante ajuda em etapas posteriores, já que o algoritmo MLP necessitava de

dados do tipo ”nominal”para um funcionamento adequado. A seguir está a lista dos atributos utilizados durante o estudo de caso:

- Tipos de Medição das Infrações - (VELOCIDADE, ALCOOLEMIA, DIMENSÃO e PESO): (Tipo Nominal)
- Nome da Marca do Veículo - (VW, TOYOTA, FIAT, GM, OUTRA, FORD, PEUGEOT, SCANIA, VOLVO, MERCEDESBENZ, RENAULT, HONDA, YAMAHA, KIA, AUDI): (Tipo Nominal)
- Turno da Infração - (MADRUGADA, MANHÃ, TARDE, NOITE): (Tipo Nominal)
- Tipo da Espécie - (CARGA, ESPECIAL, PASSAGEIRO, TRAÇÃO, MISTO): (Tipo Nominal)

3.3 CLASSIFICAÇÃO DOS DADOS DO ESTUDO DE CASO

Dentre os muitos tipos de classificadores utilizados na literatura da mineração de dados, sabe-se que os classificadores podem se comportar tanto como um conjunto de regras, uma arvore de decisão, uma rede neural, ou seja, atuar de várias maneiras para que os dados sejam classificados de forma eficiente.

Para a análise executada, foi usado o classificador Multi Layer Perceptron, usando validação cruzada (Cross-Validation), 10 Folds, devido à sua alta taxa de aprendizagem para dados não muito grandes ser satisfatória e por possui alto nível de acurácia quando trabalhado com datasets grandes, pois possui um bom desempenho (em média de 80%) para dados balanceados e bem distribuídos em vários atributos.

As seções seguintes apresentam resultados obtidos da classificação usando MLP para o Tipo, Marca, Turno e Espécie de Infrações.

3.3.1 Classificação por Tipo da Infração

O MLP classifica de forma correta, e com percentual satisfatório (mais de 80%) os tipos de infrações do estudo de caso. Obteve um resultado de 85,36% de classificações de instâncias de forma correta, equivalente à 1178 infrações, e uma taxa de erro de 14,63%, equivalente à 202 infrações; dispostas da seguinte forma:

680 infrações (Corretamente classificadas como Excesso de Velocidade); 44 infrações (Erroneamente classificadas como Alcoolemia); 9 infrações (Erroneamente classificadas como Excesso de Dimensão); 10 infrações (Erroneamente classificadas como Excesso de Peso).

Também conseguiu classificar 405 infrações (Corretamente Classificadas como Alcoolemia); 61 infrações (Erroneamente classificadas como Excesso de Velocidade); 6 infrações (Erroneamente classificadas como Excesso de Dimensão); 11 infrações (Erroneamente classificadas como Excesso de Peso).

Ainda mais, conseguiu classificar 19 infrações (Corretamente Classificadas como Excesso de Peso); 11 infrações (Erroneamente classificadas como Excesso de Velocidade); 6 infrações (Erroneamente classificadas como Excesso de Alcoolemia); 7 infrações (Erroneamente classificadas como Excesso de Dimensão).

E por último, conseguiu classificar 74 infrações (Corretamente Classificadas como Excesso de Dimensão); 15 infrações (Erroneamente classificadas como Excesso de Velocidade); 5 infrações (Erroneamente classificadas como Excesso de Alcoolemia); 17 infrações (Erroneamente classificadas como Excesso de Peso); Como demonstrado na Figura 5.

```
Correctly Classified Instances        1178   85.3623 %
Incorrectly Classified Instances       202   14.6377 %
Total Number of Instances             1380
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 680  44   9  10 |   a = VELOCIDADE
  61 405   6  11 |   b = ALCOOLEMIA
  11   6  19   7 |   c = DIMENSAO
  15   5  17  74 |   d = PESO
```

Figura 5 – Classificação de Infrações Por Tipo de Infração

3.3.2 Classificação por Marca Do Automóvel

Para a Marca do Automóvel, o classificador teve uma grande dificuldade de classificar de forma correta, devido ao grande número de atributos, obtendo um resultado não muito satisfatório de 25% de acerto, equivalente à 345 infrações. Para a marca VW : 91 acertos e 155 erros; TOYOTA : 0 acertos e 30

erros; FIAT : 10 acertos e 153 acertos ; GM :130 acertos e 122 erros; e assim sucessivamente para as outras marcas, mantendo o baixo percentual de acerto para esse atributo, como demonstrado na Figura 6.

```
Correctly Classified Instances         345 25      %
Incorrectly Classified Instances      1035 75      %
Total Number of Instances             1380
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
  91   1  10 103  20  21   0   2 |   a = VW
   2   0   0  13  11   4   0   0 |   b = TOYOTA
  57   1  10  80  12   3   0   0 |   c = FIAT
  78   3   8 130  24   9   0   0 |   d = GM
  40   6   2  44  50  18   0   4 |   e = OUTRA
  48   2   5  44  28  23   0   0 |   f = FORD
  10   0   3  16   0   1   0   0 |   g = PEUGEOT
   0   0   0   0   6   0   0  17 |   h = SCANIA
```

Figura 6 – Classificação de Infrações Por Marca de Automóvel

3.3.3 Classificação por Turno da Infração

Em relação ao turno, o resultado foi de percentual de 95.21%! Totalizando o acerto de 1314 infrações de forma correta, tendo apenas 4,78% de taxa de erro, equivalente à 66 infrações classificadas de forma errada. Para a infração no turno da Madrugada, o MLP obteve 208 acertos e apenas 3 erros. Para o Turno da Manhã 475 acertos e 31 erros; Turno da Tarde 431 acertos e 18 erros e o turno da Noite 200 acertos e 14 erros, como demonstrado na Figura 7.

```
Correctly Classified Instances        1314   95.2174 %
Incorrectly Classified Instances        66    4.7826 %
Total Number of Instances             1380
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 208   2   1   0 |   a = MADRUGADA
  15 475  16   0 |   b = MANHA
   0  12 431   6 |   c = TARDE
   0   0  14 200 |   d = NOITE
```

Figura 7 – Classificação de Infrações Por Turno da Infração

3.3.4 Classificação por Espécie de Infração

Já para a Espécie de Infração (Carga, Especial, Passageiro, Tração e Misto), o Classificador MLP teve uma pequena queda de percentual de acerto em relação

ao Turno, porém ainda conseguiu se manter positivo em relação à diferença entre acertos e erros, tendo 73,98% de acertos, equivalentes à 1021 infrações; e 26,01% de erros, equivalentes à 359 infrações classificadas de forma errada. Sendo 106 acertos e 140 erros para Carga; 5 acertos e 53 erros para Especial (por possuir poucos casos do tipo Especial, o classificador teve dificuldade de classificar corretamente); 838 acertos e 50 erros para Passageiro (o caso mais abundante nesse atributo à ser classificado, totalizando 64,34%); 64 acertos e 25 erros para a Tração e finalmente 8 acertos e 91 erros para o tipo Misto, como demonstrado na Figura 8.

```
Correctly Classified Instances        1021  73.9855 %
Incorrectly Classified Instances       359  26.0145 %
Total Number of Instances             1380
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 106   1 120  11   8 |   a = CARGA
   6   5  44   0   3 |   b = ESPECIAL
  27   2 838   1  20 |   c = PASSAGEIRO
  15   0   9  64   1 |   d = TRACAO
   7   4  80   0   8 |   e = MISTO
```

Figura 8 – Classificação de Infrações Por Espécie de Infração

3.4 SUGESTÕES DE APLICAÇÕES DOS RESULTADOS ENCONTRADOS

Com base nos resultados encontrados da classificação dos dados para esse domínio especifico de infrações de trânsito, adquirimos informações relevantes para a aplicação delas no dia-a-dia. Como já citado neste trabalho, a BR-116 é extremamente movimentada, possuindo um alto índice de infrações, e isso foi bem mais evidenciado com a classificação das 1380 instâncias pelo classificador MLP, pois pra cada atributo (Tipo, Marca, Turno e Espécie) o classificador conseguiu extrair informações que podem ser ricamente acrescentadas às autoridades de trânsito responsáveis, tanto em âmbito municipal, estadual e federal, com o intuito de aprimorar a fiscalização de trânsito ,e até mesmo, se possível, prevenir a ocorrência dessas infrações.

Com os dados provenientes dessa análise, vimos a real importância da extração de dados relacionados a infrações de trânsito, pois os dados coletados também servem de impulso de desenvolvimento de novas aplicações

computacionais relacionadas ao trânsito, apoiando os órgãos reguladores (CONTRANs) e fiscalizadores de trânsito (DETRANs) para uma melhoria contínua e significativa de um trânsito mais saudável.

Como exemplo disso, temos que a maioria (743 infrações) das infrações do quesito Tipo de Infração catalogadas pelo classificador é composta de Excesso de Velocidade, como demonstrado no Gráfico 1.

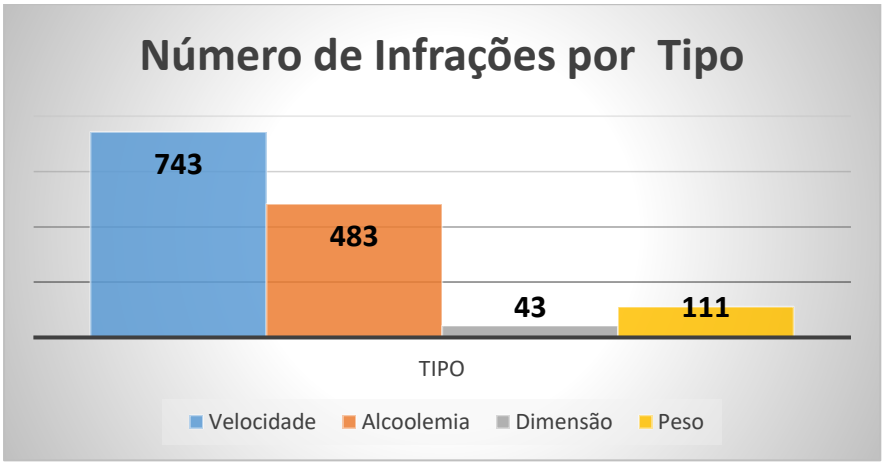


Gráfico 1 – Número de Infrações por Tipo

Isso direciona as autoridades fiscalizadoras a investirem mais em semáforos, placas reguladoras de velocidade e o aumento de contingente de agentes de trânsito nesses trechos de alto índice de infrações desse tipo.

Também vimos que o classificador conseguiu extrair dados relacionados ao Turno da infração, nos mostrando que o horário mais incidente de infrações (508 infrações) é no turno matinal, como demonstrado no Gráfico 2, e a Espécie que mais comete infrações é do tipo Passageiro (888 Infrações), como demonstrado no Gráfico 2.

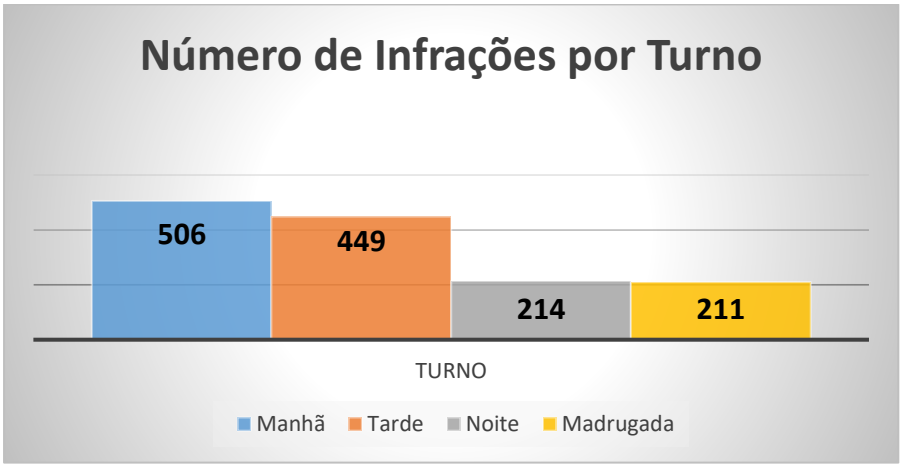


Gráfico 2 – Número de Infrações por Turno

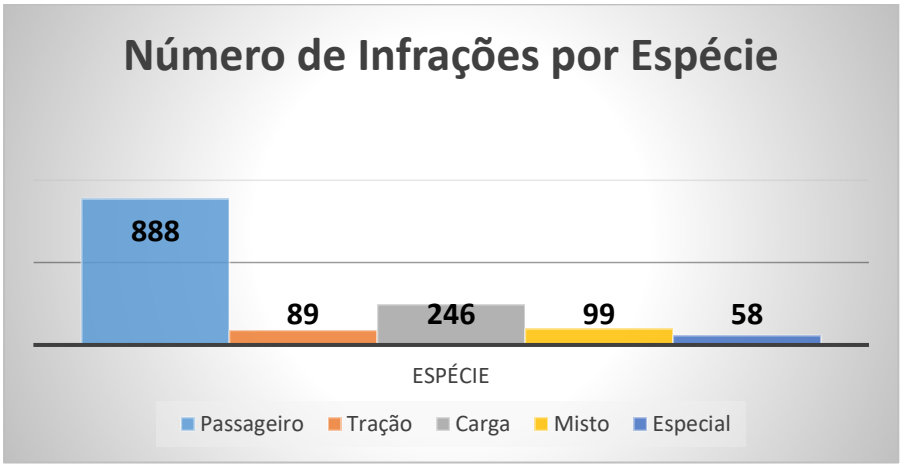


Gráfico 3 – Número de Infrações por Espécie

Isso instiga às autoridades de trânsito a saber o real motivo dessas infrações ocorrerem nesse turno, e da maior parte ser da espécie passageiro, seja por desatenção devido ao celular no volante; o atraso para o trabalho; stress acumulado e assim prover soluções adequadas para melhorias no trânsito através de fiscalização mais acentuada e por campanhas de conscientização aos condutores de veículos terrestres.

4 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho explicitou um estudo de caso sobre infrações de trânsito em vias terrestres, mostrando o conceito de trânsito; índices de infrações no período 2007 a 2013 no Rio Grande do Sul; o conceito de mineração de dados e o classificador utilizado.

No Capítulo 2 mostrou-se o conceito teórico sobre Mineração de Dados, suas ferramentas disponíveis, e suas principais funções de aplicação nos mais diversos campos através da classificação de dados e foi abordado o classificador *Multi Layer Perceptron* .

Já no Capítulo 3 foi feita a análise do estudo de caso de ocorrências de dados de trânsito em vias terrestres e suas formas de avaliação através de taxa de erro e matriz de confusão. Foi explicitada a descrição dos dados utilizados, através do dataset adquirido da PRF[1], também foi apresentada a definição do dataset em termos de objetos e atributos utilizadas de forma computacional.

Logo em seguida, foi feita a classificação de dados de forma concisa através do classificador MLP através dos atributos escolhidos: Tipo da Infração, Marca do Automóvel, Turno da Infração, Espécie da Infração.

Na Seção 3.4, vimos que as informações encontradas por meio da classificação de dados de infrações em vias terrestres são de grande ajuda para a prevenção e fiscalização dessas infrações na BR-116, pois por meio das instâncias classificadas conseguimos extrair padrões de ocorrências de trânsito.

Este trabalho deixa sua contribuição com as informações adquiridas por meio deste estudo de caso aplicando a classificação de dados, pois vimos a real importância dos dados relacionados à infração de trânsito, pois das 1380 instâncias classificadas, temos que 743 Infrações são do tipo Excesso de Velocidade; 508 Infrações foram cometidas no turno da Manhã e 888 Infrações foram cometidas pela Espécie do Tipo Passageiro. Essas informações

[1] http://dados.gov.br/dataset/multas-rodovias-federais

Servem como incentivo para o desenvolvimento de aplicações práticas para fiscalização e prevenção de infrações de trânsito.

Para trabalhos futuros, temos a ampliação do uso de infrações de trânsito existentes, totalizando mais de 200 tipos, pois neste trabalho foram expostos apenas 4 tipos, (Excesso de Velocidade, Alcoolemia, Excesso de Dimensão e Excesso de Peso) a possível implementação de um software capaz de reconhecer padrões de infrações de trânsito em todo o território nacional. Uma iniciativa como essa, ao automatizar a fiscalização de trânsito, tem o potencial de amenizar os impactos causados pelas infrações de trânsito, salvando vidas e evitando gastas em escala nacional.

REFÊRENCIAS

AMO, S.de.Técnicas de mineração de dados. **Jornada de Atualização em Informática, 2004.**

BARBON,**Barbon**.2018.<http://www.barbon.com.br/wpcontent/uploads/2016/04/FundamentosInteligenciaArtificial-3.pdf> acessado em setembro de 2018.

BARBON,**Barbon**.2018.<http://www.barbon.com.br/wpcontent/uploads/2016/04/FundamentosInteligenciaArtificial10.pdf> acessado em setembro de 2018.

BATISTA,P.R.L. **Data mining na identificação de atributos valorativos da habitação**. Dissertação (Mestrado)—UniversidadedeAveiro,2010.

CAMILO, C.O.;SILVA, J. C .d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás(UFC)**, p.1–29,2009

CÔRTES,S. da C .;PORCARO,R.M.;LIFSCHITZ,S. **Mineração dedados-funcionalidades, técnicas e abordagens**. [S.l.]:PUC,2002.

DEAMO,**Deamo**.2018<http://www.deamo.prof.ufu.br/arquivos/Aula2.pdf> acessado em agosto de 2018.

DETRAN-RS,**detran-rs**<http://www.detran.rs.gov.br/conteudo/29998/perigo-nas-estradas%3A-infratores-sao-homens,-entre-26-e-30-anos>

DIAS, M.M.;FILHO,L.A. da S.;LINO,A.D.P.;FAVERO,E.L.;RAMOS,E.M.L.S. Aplicação de técnicas de mineração de dados no processo de aprendizagem na educação a distância. In: **Brazilian Symposiumon Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.:s.n.],2008.v.1,n.1,p.105–114.

HAN, J.;KAMBER,M. **Data Mining:Concepts and Techniques**, **University of Illinois at Urbana-Champaign.** [S.l.]:Elsevier,2006.

IAEXPERT,**Iaexpert**.2018. <http://iaexpert.com.br/index.php/2016/12/29/classificador-zeror-no-weka/>n acessado em setembro de 2018.

JUSBRASIL,**jusbrasil** <https://www.jusbrasil.com.br/topicos/10632340/artigo-1-da-lei-n-9503-de-23-de-setembro-de-1997>

MVFIDELIS,**Mvfidelis**.2018.<https://pt.slideshare.net/mvfidelis/construcaode-classificadores> acessado em setembro de 2018.

PANG-NING, T. ; STEINBACH, M .;KUMAR,V. Introdução ao “datamining”. **Rio de Janeiro: Ciência Moderna**, 2009.

UFLA,**Ufla**.2018.<http://repositorio.ufla.br/bitstream/1/5359/1/MONOGRAFIA_Treinamento_de_redes_neurais_artificiais_utilizando_algoritmos_geneticos_em_plataforma_distribuida.pdf:> acessado em setembro de 2018.