UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

KAIO RIBEIRO ANDRIANI

APLICAÇÃO DE TÉNICAS DE APRENDIZADO DE MÁQUINA NO SETOR DE RESSARCIMENTO DE UMA EMPRESA SECURITÁRIA

LONDRINA

2021

KAIO RIBEIRO ANDRIANI

APLICAÇÃO DE TÉNICAS DE APRENDIZADO DE MÁQUINA NO SETOR DE RESSARCIMENTO DE UMA EMPRESA SECURITÁRIA

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Engenharia de Produção, do Departamento de Engenharia de Produção, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Rogério Tondato

LONDRINA

2021

KAIO RIBEIRO ANDRIANI

APLICAÇÃO DE TÉNICAS DE APRENDIZADO DE MÁQUINA NO SETOR DE RESSARCIMENTO DE UMA EMPRESA SECURITÁRIA

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 16/agosto/2021

_____

José Ângelo Ferreira
Doutor
Universidade Tecnológica Federal do Paraná

_____

Silvana Rodrigues Quintilhano
Doutora
Universidade Tecnológica Federal do Paraná

_____

Rogério Tondato
Doutor
Universidade Tecnológica Federal do Paraná

LONDRINA

2021

RESUMO

Diante de variáveis que afetam o processo de uma empresa, às vezes, resolver os problemas pelo método cartesiano torna-se muito oneroso. Neste sentido, a modelagem e simulação, em especial os algoritmos de aprendizado de máquina, podem contribuir para a melhoria dos processos de fabricação. Assim, o objetivo deste trabalho, foi testar algoritmos de aprendizado de máquina para processos judiciais de cobrança de uma seguradora, tentando prever o resultado da ação. O método utilizado neste trabalho foi a Modelagem e Simulação, onde foi possível utilizar os algoritmos de Redes Neurais, Regressão Logística e Árvores de Decisão. Os resultados alcançados chegaram a quase 80% de predição, indicando uma adequada escolha dos preditores. Com esta pesquisa, foi possível demonstrar a possibilidade de predição através dos algoritmos e assim, possuir mais uma ferramenta para a tomada de decisão durante a consecução dos processos da seguradora.

Palavras-Chave: Algoritmos, Redes Neurais, Seguradora.

ABSTRACT

Faced with variables that affect a company's process, sometimes, solving problems using the Cartesian method becomes very costly. In this sense, modeling and simulation, especially machine learning algorithms, can contribute to the improvement of manufacturing processes. Thus, the objective of this work was to test machine learning algorithms for legal collection proceedings of an insurance company, trying to predict the outcome of the action. The method used in this work was Modeling and Simulation, where it was possible to use the algorithms of Neural Networks, Logistic Regression and Decision Trees. The results achieved reached almost 80% prediction, indicating an adequate choice of predictors. With this research, it was possible to demonstrate the possibility of prediction through algorithms and thus, have one more tool for decision making during the achievement of the insurer's processes.

Keywords: *Algorithms, Neural Networks, Insurance.*

SUMÁRIO

1. INTRODUÇÃO

As novas tecnologias de informação estão revolucionando continuamente a forma de lidar com a comunicação e armazenamento de dados. O avanço da tecnologia traz a possibilidade de medir e quantificar ações, tomada de decisões e resultados empresariais, fazendo com que a análise desses dados gerados se torne crucial para o desenvolvimento empresarial. Vale notar que quantidade de dados gerados pode ser tão alta, que a maior parte das empresas não consiga usufruir dos benefícios de possuir todos esses dados.

Nesse sentido, torna-se necessário um investimento de recursos para examinar os dados e transformá-los em informações. Kotler (2000), demonstra a importância da informação, afirmando que ela é vendida e comercializada como um produto, seja ao pagar para realizar algum tipo de curso ou ao comprar revistas ou livros. “A produção, a embalagem e a distribuição de informações constituem um dos principais setores econômicos da sociedade de hoje” (Kotler, 2000).

Para se obter uma gestão eficiente é necessário o planejamento ter como base informações precisas, para que dessa forma, a tomada de decisão se torne mais concisa. Logo, metrificar e possuir informações concretas sobre o resultado de cada ação da empresa se torna crucial em qualquer tipo de setor.

Com o objetivo de trabalhar a massiva quantidade de dados gerados fazendo com que a máquina aponte qual pode ser a melhor escolha para uma tomada de decisão. Técnicas de aprendizado de máquina têm sido utilizados com sucesso para esse objetivo. Algoritmos como as redes neurais, regressão logística e árvore de decisões são exemplos que serão citadas nessa pesquisa.

Visto que não existe um algoritmo pré-definido que funcione de maneira ideal para cada tipo de problema, se torna necessário entender a fundo os detalhes do problema tratado e por muitas vezes, realizar testes com diferentes tipos de algoritmos para comparar os resultados encontrados.

No caso de uma empresa seguradora, mais especificamente no setor de ressarcimento, cujo um dos objetivos é reaver um valor gasto através de processos judiciais, a confiabilidade e quantidade dos dados trazem mais precisão e

confiabilidade na tomada de decisões. Nesse sentido, um estudo para analisar e avaliar a chance real de um processo judicial ser ou não capaz de ressarcir um determinado valor, pode se tornar interessante ao trazer resultados que impactam diretamente no resultado do setor.

A principal problemática deste trabalho, se enquadra na análise probabilística de determinada ação judicial, ser cabível de ter o valor gasto da empresa B, estornado ou não. Problemas para obter essa resolução, também podem ser encontrados, como a falta de coleta de dados ou escassez de dados para uma análise precisa, devida ao fato da empresa B, possuir um histórico curto de coleta de dados, com esse intuito.

Serão utilizadas técnicas que são consideradas relativamente novas para o tratamento dos dados da empresa. São chamadas comumente de mineração de dados, ou do inglês, *data minin*g. E como ferramentas principais estão inclusas: redes neurais, regressão logísticas, árvores de decisão, vetores de suportes, entre outros citados ao longo do trabalho (GARGANO & RAGGAD, 1999).

1.1 OBJETIVO

Essa pesquisa tem como objetivo geral aplicar algoritmos de aprendizado de máquina para buscar a probabilidade de um processo judicial a ser ressarcido da empresa B. Como objetivos específicos tem-se:

- Definir os métodos de aprendizado de máquina que possibilitem a verificação e a viabilidade da judicialização;
- Otimizar as tomadas de decisão para judicialização de um processo;
- Diminuição de custos desnecessários com judicialização de processos;
- Analisar os resultados obtidos.

1.2. JUSTIFICATIVA

Atualmente, as empresas lidam com um fator inédito: a altíssima quantidade de dados. Com a competividade do mercado e a necessidade constante de redução de custos, definir a melhor estratégia de vendas, tendo como base dados e análises concretas para essa decisão, se torna crucial para decisões acertadas, principalmente

quando o objetivo é concentrar os esforços em apenas algumas opções dentro de um universo de escolhas.

Essa pesquisa faz-se necessária para identificar os processos judiciais que possuem maiores chance de se obter o valor estornado para empresa, levando em consideração dados como valor do processo, localidade geográfica do terceiro, do fórum do processo, quantidade de movimentações do processo, entre outros.

Essa pesquisa também impacta positivamente a carreira profissional do pesquisador, pois aumenta a gama de oportunidades dentro da área de ciência de dados e afeta de maneira positiva a imagem dentro da companhia.

Assim essa pesquisa contribuirá para dar parâmetros de resultados para empresa em análise e para empresas que também possuem o processo de ressarcimento de valores.

2. REFERENCIAL TEÓRICO

Para alcançar e abordar o tema de máquina de aprendizado, se faz necessário comentar brevemente sobre mineração de dados e análise de dados, visto que são essenciais para a aplicar os modelos de aprendizado de máquina.

O aprendizado de máquina está dentro da área da inteligência artificial. A inteligência artificial utiliza a técnica de obter conhecimento através de amostra de dados, ou seja, simula o ato humano de adquirir habilidade através da experiencia. Para cumprir esse objetivo, são utilizados uma série de algoritmos que possuem a capacidade de classificar conjuntos de elementos. Entende-se que classificação, é o processo de atribuir a um dado uma variável resposta, ou um rótulo a qual ele pertence. Portanto, dado um conjunto de dados, a máquina, deve ser capaz de rotular corretamente cada conjunto (REZENDE et al., 2003).

Para Lorena (2007) o aprendizado de máquina, é uma técnica que é aplicada um princípio de inferência denominado indução, logo pode ser dividido de duas formas, supervisionado e não supervisionado.

No aprendizado supervisionado, um agente externo fornece a máquina os rótulos corretos para os conjuntos, ou seja, nesse tipo de aprendizado é fornecido a variável resposta correta, e então o algoritmo se ajusta baseando-se nos rótulos fornecidos (SANTOS, 2019).

Já no aprendizado não supervisionado, não existe a presença do agente externo fornecendo a variável resposta correta, nesse método, o algoritmo absorve e processa as entradas, analisa as saídas e o seus padrões, e classifica automaticamente (FERNEDA, 2006).

Sendo muito utilizado, a área de aprendizagem de máquina possui métodos que mesmo dado um pequeno conjunto de dados, é feito a busca por padrões, o que possibilita a classificação de dados. Essa abordagem, pode ser utilizada para realizar predições de diversas situações, como por exemplo prever diagnósticos (PRATI e CRISTIANO, 2006).

A análise prescritiva, de acordo com Blanchard e Morison (2013), são a utilização de padrões que foram encontrados no passado, para sinalizar tendências

futuras. São coletados dados de anos anteriores da empresa, geralmente dando foco em uma área de negócio da organização ou em uma determinada ação específica.

De acordo com Silveira (2013), com a alta quantidade de dados gerados por uma empresa, e o ambiente corporativo vem se mostrando cada vez mais competitivo, e assim a análise de dados se torna uma ferramenta crucial, para o processo decisório do médio e alto gerenciamento, pois subsidia com informação e dá suporte a tomada de decisão. Carvalho (2019), ilustra a enorme quantia de dados gerados por dia; o projeto Genoma, possui bilhões de bases genéticas, armazenando, milhares de dados para cada uma dessas bases; ou então, instituições que mantêm repositórios com milhares de transações dos seus clientes.

Essa alta densidade de dados ocasiona com que, as técnicas tradicionais de tratamento de dados, se tornem ultrapassadas, principalmente devido a essa alta quantidade de informação. Com o objetivo de suprir essa necessidade, no final da década de 80, surge o *Data Mining*, traduzido para o português como Mineração de dados (FREITAS, 2019).

Lorose (2005), afirma que as empresas investem um alto capital na coleta de dados, e muitas vezes nenhuma informação útil é identificada, fazendo com que a Mineração de Dados, se torne tão importante e promissora. Não somente pela iniciativa privada, afirma Chakrabarti (2006), mas o setor público e o terceiro setor (ONGt's) também podem usufruir do *Data Mining*, porém corre o risco de serem ricos em dados, mas pobre em informação.

Para Camilo e Silva (2009), a Mineração de Dados é um assunto considerado multidisciplinar, e varia de acordo com o autor ou com o campo de aplicação, contudo, de forma geral, há três áreas que são mais utilizadas e representativas no *Data Mining*, que seriam: A estatística, o aprendizado de máquina (do inglês *Machine Learning*) e o banco de dados.

Para Hand et al. (2007), com uma visão estatística, a "Mineração de dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados".

Abordando por uma perspectiva de banco de dados, Camilo (2010), afirma que a Mineração de Dados, busca não só dados que sejam utilizáveis, mas também compreensível ao usuário. Ele também diz que “Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização”.

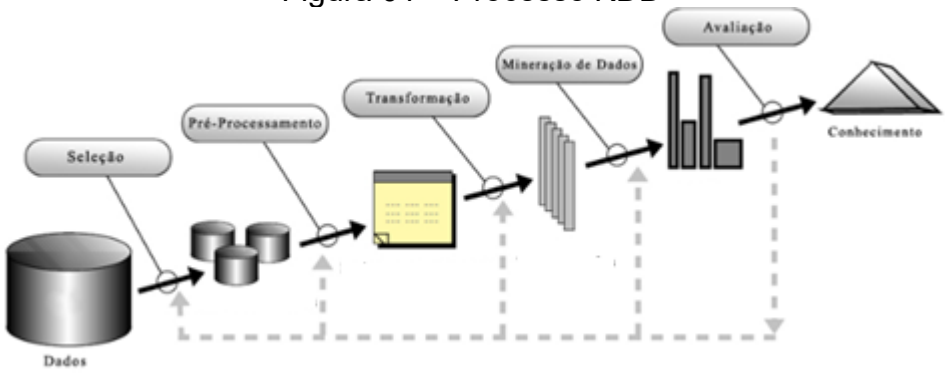
E por fim, de uma perspectiva de banco de dados, Bernardi (2010), afirma que a mineração de dados ainda é limitada por fatores computacionais, mesmo utilizando certos algoritmos de descoberta e ainda é considerado um dos passos para a descoberta de conhecimento dentro de um tópico.

Com o objetivo de elucidar o foco da Mineração de dados, Witten et al. (2000), expõe exemplo de setores que o *Data Mining*, é mais efetivo: Em bancos ou serviços de “*Telemarking*”, para a identificando padrões no comportamento do cliente para auxiliar no seu tratamento, ou melhoria no acesso a sua informação; No setor da saúde, aumentando a precisão de diagnósticos; Na área de segurança digital, identificando padrões de comportamento específicos, para antecipar ataques terroristas.

Ponniah (2001) também cita áreas em que a Mineração de dados pode trazer melhorias, como por exemplo: A Logística, com aprimoramento nas escolhas de rotas e distribuição de produtos; No Marketing, melhoria no direcionamento da publicidade; Vendas, identificação do padrão de consumo dos clientes.

Para Santos et al. (2007), o processo de Mineração de Dados é uma das etapas de outro processo, nomeado de KDD (*Knowledge Discovery in Databases*). Segundo Santos et al (2007), O KDD, surge com a mesma perspectiva do *Data Mining*, suprir a necessidade que a era da informação trouxe: Tratar de forma efetiva, a grande quantidade de dados gerados pela empresa. Ainda para o mesmo autor, o KDD engloba a Mineração de Dados, pois ele seria todo processo que transforma os dados em conhecimento, um processo realizado de acordo com a figura 01.

Figura 01 – Processo KDD



Fonte: Adaptado de Camilo e Silva (2009)

Storopoli (2016) define o KDD, como um processo complexo, que busca encontrar novos padrões, que sejam pertinentes e proveitosos para determinada ações da empresa.

Silveira (2013), diz que geralmente, essa alta quantidade de informação fica armazenada de formas diferenciadas entre si, como em banco de dados, documentos ou arquivos. Com o objetivo de unificar essa informação em um apenas um local, surgem o *Data Warehouse* (DW).

Ainda Silveira (2013), para transformar essa alta quantidade de dados em informação e sob a mesma ótica do *Data Warehouse*, as aplicações OLAP (*Online Analytical Processing*), são ferramentas, com o objetivo de analisar essa alta quantia de dados para determinado objetivo, conectam-se com o DW e tornam possível aplicações de diversas funções de análises. Logo, o OLAP, além de gerar uma interface amigável, o usuário (ainda que geralmente tenha a necessidade de ser capacitado) é capaz de manipular os dados para gerar relatórios analíticos, tabelas pivô, gráficos, entre outros.

De acordo com Turban, Sharda e Arason (2009), tanto o DW e o OLAP, são ferramentas tecnológicas conhecidas como *Business Intelligence* (BI), que abordando de forma superficial, tem como o objetivo, já citado, de colher dados e organizar, para suporta a algum tipo de decisão.

Cabe introduzir brevemente o conceito principal de cada uma das principais tarefas que a mineração de dados pode realizar, visto que será utilizado como ferramenta de análise dos dados coletados.

De acordo com Camilo e Silva (2009), a tarefa de classificação é considerada uma das mais comuns, ela tem utiliza de dados prévios de determinado objeto e então o classifica, alocando-o em um grupo. Por exemplo, tendo como base de dados informações de várias classes de uma escola, com a classificação, é possível determinar para qual classe um aluno novo deverá ser alocado.

A tarefa de regressão ou estimação, é semelhante à de classificação, contudo ao invés de utilizar valores categóricos, utiliza valores numéricos. Ela, assim como a de classificação, absorve dados de determinado objeto já classificados e perante a um novo objeto, é capaz de classificá-lo, porém na regressão, de forma numérica. Por exemplo, ao absorver dados de gastos médios de várias famílias no período de início do ano letivo dos filhos, ela é capaz de obter uma quantia numérica de qual irá ser o gasto de outra família (CAMILO; SILVA, 2009).

A tarefa de agrupamento ou do inglês, *clustering*, tem como objetivo agrupar objetos com traços em comum, facilitando a visualização de padrões ou nichos. Como diferenciação das anteriores, ela não necessita de dados já classificados, visto que ela não tem como resultado final, predizer ou estimar o valor de uma variável. Ela pode ser útil para segmentar um nicho de produto ou identificar comportamentos humanos específicos que resultam em ações indesejadas (CAMILO; SILVA, 2009).

Ainda de acordo com Camilo e Silva (2009) a tarefa de associação é uma das que obtém melhores resultados, ela utiliza o a lógica condicional, identificando quais objetos estão relacionados entre si. Esse tipo de tarefa é muito utilizado para encontrar produtos do mercado que são vendidos em conjunto como o a carne e a cerveja. Também pode ser utilizado para encontrar efeitos colaterais causados por medicamentos ou perfis de usuários que correspondem positivamente a um tipo oferta.

2.1 MODELOS DE CLASSIFICAÇÃO

2.1.1 Baesiyano

De acordo com Kinas e Andrade (2017) afirmam que para realizar uma análise estatística, geralmente são utilizados dois métodos, o convencional ou o bayesiano. O modelo convencional é de fato o mais utilizado, principalmente por motivos históricos e por obstáculos computacionais. Contudo, essa soberania tende a mudar,

devido as vantagens claras do modelo bayesiano, como por exemplo Kinas e Andrade (2017) citam “o modelo bayesiano não precisa necessariamente estar associada a fenômenos medidos por frequência relativa ou então a maior proximidade do conceito popular de probabilidade, pois colhe a probabilidade como uma medida racional e condicional de incerteza.”

De acordo com Souza (2016), a principal característica da inferência bayesiana é que ela aborda um parâmetro θ de uma distribuição *f(x|θ)* como uma variável aleatória. Ou seja, é um modelo que considera o quanto o pesquisador conhece sobre o experimento, antes de realizar a amostra.

Para Ehlers (2003) a ideia do teorema de *bays*, consiste em que o pesquisador desconhece sobre a variável, ou seja, o valor de *θ* é desconhecido e o objetivo é justamente, diminuir esse fator. É intuitivo que, conforme o observa-se em uma quantidade aleatória *X*, o desconhecimento de *θ* tende a diminuir, sendo que ele está resumido probabilisticamente através de *p(θ)*, de acordo com a figura 02.

Figura 02 – Fórmula bayesiana

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d\theta}.$$

Fonte: Própria (2019)

1/p(x), não depende de θ, ele é uma constante normalizadora de p(θ|x).

A forma usual está representada na figura 03.

Figura 03 – Fórmula bayesiana usual

$$p(\theta|x) \propto l(\theta; x)p(\theta).$$

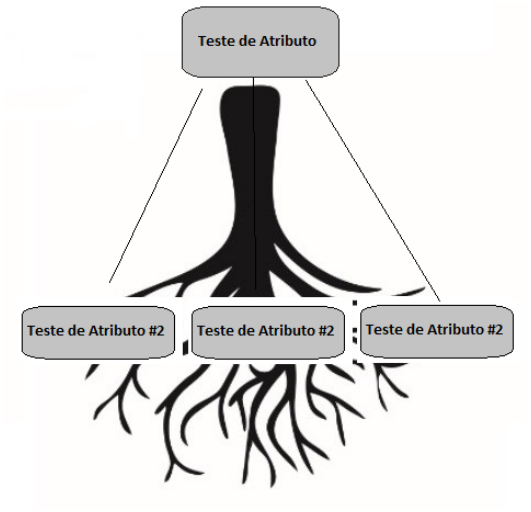
Fonte: Própria (2019)

Cujo escrita seria: distribuição a posteriori ∝ verossimilhança × distribuição a priori.

2.1.2 Árvore de Decisão

A árvore de Decisão se inclui em um método utilizado para a classificação de dados, segundo Breiman (1984), é um modelo, que metaforicamente, representa uma árvore invertida, com “nós e ramos”. O primeiro nó, seria a raiz, os seguintes são formados através de decisões baseadas no nó raiz. Cada nó seguinte do nó raiz, é uma avaliação e resolução do nó raiz, formando assim os ramos. No fim da Árvore de Decisão estão as folhas da árvore, que representa a previsão ou resolução da raiz, conforme a figura 04.

Figura 04 – Árvore de decisão



Fonte: Própria (2019)

Para Gama (2000), a Árvore de Decisão, utiliza de um método comum e tático para a resolução de problemas. Ela subdivide esse problema em outros menores, para assim ser possível tratar cada um deles individualmente e com melhor foco.

Quando há necessidade de visualizar e representar os resultados de forma hierárquica, o modelo da árvore de decisão, é o único capaz de trazer esse benefício. Com o problema dividido e ranqueado de forma estruturada, a tomada de decisão tende a dar maior consideração para os problemas cujo seriam os de maior importância, os do topo (COLARES, 2010).

A Árvore da Decisão se torna uma ilustração de fácil compreensão para a base de código computacional. Onde, por exemplo, cada ramo se torna uma regra do tipo

“se-então” (INGARGIOLA, 1996). Witten e Frank (2000) afirmam que atualmente, há uma grande quantidade de algoritmos que utilizam a Árvore da Decisão como base.

Com o intuito de evitar erros e avaliar a qualidade da Árvore de Decisões, Bradzil (1999), afirma que é adequado aplicar testes com dados que não foram utilizados na criação do modelo, estimando também, a capacidade com que a Árvore de Decisão tem de se adequar a novas situações e o quanto ela é capaz de generalizar os dados.

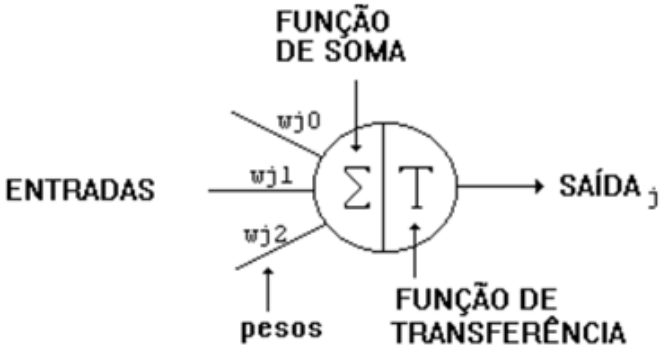
2.1.3 Redes Neurais

Ainda sobre mineração de dados, e com o objetivo de predizer de informações, Martins et al (2008), diz que Redes Neurais, procura de uma forma simples, traduzir o mesmo método que animais usam para aprender algo, utilizando a experiência, para modelos matemáticos, logo, como o próprio nome diz, é inspirado em redes neurais.

Um neurônio biológico é constituído por um núcleo, a soma, citoplasma, uma membrana celular e os caminhos de entrada e saída, os dendritos e os axônios, respectivamente. Nota-se que os axônios são geralmente mais extensos se comparados com os dendritos (MARTINS et al, 2008).

Já em uma rede neural artificial, os dendritos, assim como os neurônios biológicos, são os caminhos de entrada de dados, “os pesos” seriam as sinapses do neurônio, a “função de soma” são os estímulos que o neurônio recebe e a “função de transferência” seria o limiar de descarga, ou seja, cada elemento do neurônio biológico é representado no neurônio artificial, conforme a figura 05 (Tafner, 1998).

Figura 05 – Neurônio único



Fonte: Adaptado de Tafner (1998)

Assim como o sistema biológico, as redes neurais artificiais recebem toda a informação, de diferentes entradas, e a distribui de forma coordenada para todo resto do organismo. “Normalmente as informações armazenadas por uma RNA são distribuídas para todas as unidades de processamento, estando isso diretamente ligado ao sistema atual dos computadores, de armazenar informação em um endereço de memória” (FINOCCHIO, 2014).

As saídas de um neurônio, se tornam as entradas para outro, construindo ligações, ocorrendo a união de diversos neurônios onde se formam as sinapses e as informações são transportadas. Esse conjunto de neurônios artificiais representa uma rede neural artificial. Tendo a perspectiva como um todo em uma rede neural, cada conjunto de neurônio artificial assume uma função, ou seja, cada neurônio fica com uma parte do processo, neurônio de entradas, os neurônios intermediários e os neurônios de saída (Tafner, 1998).

Para exercer o aprendizado máquina nas redes neurais, o algoritmo trabalha alterando os pesos nos neurônios artificiais, esses pesos, já iniciam com um parâmetro definido para evitar erros ou desvio bruscos. Vale notar que os RNAS, são feitos com uma base de dados já existente, geralmente experimental e quanto maior for essa base de dados, mais “experiência” o RNA terá, logo essa quantidade de torna importante para a qualidade do resultado final (Finocchio, 2014).

Ao finalizar a construção do modelo, e caso se deseje adicionar uma nova base de dados, a rede neural irá recomeçar, ou seja, será calculado novas estáticas baseadas nos novos dados, logo a rede terá que calibrar e aprender novamente perante os novos dados. Assim que a rede neural estiver com aprendizado suficiente, ela será capaz de obter previsões dos dados de entrada (FINOCCHIO,2014).

2.1.4 Regressão Logística

A Regressão Logística é utilizada para suprir necessidades que outros métodos não conseguem obter, como por exemplo a regressão linear simples ou múltipla. Ela é utilizada, principalmente quando é necessário utilizar a variável dependente qualitativa e é expressa por duas ou mais categorias, logo, admite mais do que um único valor. A Regressão Logística busca prever a chance de determinado

acontecimento ocorrer ou não, ou seja, encontrar a probabilidade de ocorrência (Figueira, 2006).

Os valores que a variável dependente pode assumir são nominais, ordinais ou binários, sendo que para cada caso é utilizado um tipo de Regressão Logística. Quando a variável é ordinal, existe uma disposição ordenada entre os valores, então se enquadra no cenário de Regressão Logística Ordinal. Já quando não há uma ordem correta entre os valores, se enquadra em Regressão Nominal. Quando a variável resposta apresenta apenas 2 valores, então seria para a Regressão Logística Binária. Há casos em que existe mais de uma variável independente, portanto se enquadra em uma Regressão Logística Múltipla. E por fim, é cabível que, a variável dependente, com natureza nominal, possua mais de dois níveis de codificação, logo uma Regressão Logística Multinominal. (Figueira, 2006).

Camargo, Camargo e Araújo (2012), citam que Abordando a Regressão Logística de uma forma geral, a probabilidade de ocorrer um evento, de acordo com a figura 06.

Figura 06 – Regressão logística

$$\mathbf{P(Y = 1) = \frac{1}{1 + e^{-g(x)}}}$$

No qual:

$$g(x) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

Fonte: Própria (2019)

Nota- se que uma das vantagens da regressão logística é o fato de ela possuir um alto índice de generalidade. Nota-se que essa característica é reafirmada devido a sua curva no formato de letra S, variando os valor de X e com determinado valores para os coeficientes B0,B1,...,Bp. Portando quando g(x) tende ao infinito positivo(sinal), então P(Y=1) tende a 1 e quando g(x) tende ao infinito negativo (sinal), então P(Y=1) tende a 0. (Camargo, Carmargo e Araujo 2012)

2.1.5 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (*Support Vector Machines- SVMs*) foi criada em 1995 por Vapnik, ela é empregada para a classificação de dados, cujo tinha como objetivo inicial apenas a classificação de padrões linearmente separáveis. É uma técnica baseada na Teoria de Aprendizado Estatísticos (Lorena e Carvalho 2003).

A técnica Máquina de Vetores de Suporte fundamenta-se em Teorias Estatísticas e tem sido cada vai mais utilizada para a detecção de padrões (Cristianini e Shawe-Taylor, 2000). Principalmente, por se destacar com melhores resultados se comparada a algoritmos similares. Fato relatado em tópicos como, o reconhecimento facial (HEARST ET AL., 1998), em classificação de textos (HEARST ET AL., 1998) e em áreas de biologia que utilizam a aplicação de técnicas de informática (Zie net al., 2000). Já em outros, o algoritmo SVM se iguala, como com as Redes Neurais (HAYKIN, 1999).

A técnica de Máquina de Vetores de Suporte, se destaca quando é utilizada com o foco em algumas características, como: A possibilidade de ótima qualidade de generalização, os classificadores encontrados pela SVM, geralmente atingem ótimos resultados de qualificação, ou seja, independente dos dados utilizados o SVM encontra bons preditores ou então na Convexidade da função objetivo, o SVM possui apenas um mínimo global, uma vantagem quando comparada com Redes neurais, que possui mínimos locais que necessitam ser minimizados (SMOLA *et al.*, 1999).

O SVM separa e classifica, agrupando as variáveis em um único hiperplano, quando bem otimizado e oferecido uma base de dados, cuja não a necessidade de ser grande, o SVM estabelece uma linha de separação clara entre as duas classes. Os objetos, chamados de vetores de suporte (*Suport Vector – SV*) são os que se localizam na linha da fronteira entre a separação entre as classes e a segunda classe. “A solução para o problema de classificação é representada pelos vetores de suporte, os quais são fundamentais na determinação do hiperplano de separação com margem máxima.” (SOARES, 2008).

2.2 MODELO DE REGRESSÃO

2.2.1 Linear e Não Linear

A regressão linear tem como principal objetivo encontrar previsões, ela pesquisa numericamente, a relação entre variáveis. Uma das suas principais formas de uso, é de encontrar como o valor do dólar várias de acordo como crescimento econômico do Brasil, por exemplo. Vale notar que, para que a regressão linear seja aplicável, é necessário que haja uma relação entre as duas variáveis (PETENATE, 2019).

Para facilitar o entendimento do conceito de regressão linear e não linear, primeiramente cabe explicar a diferença entre os dois modelos. Para Thomas (2016), de forma simplificada, a principal diferença é que no modelo não lineares, são caracterizados por funções, já nos modelos lineares são compostos por variáveis dependentes, portanto nos modelos não lineares, é preciso que pelo uma das variáveis dependa de um dos parâmetros.

Para Zeviani, Júnior e Bonat (2013): “O modelo estatístico é linear se a quantidade de interesse, geralmente a média de Y, é função linear dos parâmetros, caso contrário é não linear”. Também afirmam que a motivação para optar por um modelo não linear geralmente não é empírica e que possui vantagens sobre a regressão linear como poderem serem parasimoniosos, já que possuem menos parâmetros ou então por serem baseadas em teorias ou princípios físicos, químicos ou biológicos. Contudo, também possui uma desvantagem crítica, por necessitarem de procedimentos iterativos e de valores iniciais para os parâmetros.

Os modelos não lineares são utilizados em diversos campos, como por exemplo para determinar o crescimento de vegetais com base na quantidade de nutrientes que ele recebe, ou então em o quão efetivo a aplicação de catalizadores são, em reações químicas (THOMAS, 2016).

2.2.2. Modelo de Regressão Florestas Aleatórias

O modelo de regressão de florestas aleatórias, ou do inglês, *Randon Forest*, é um algoritmo de aprendizagem que se demonstrou extremamente eficaz e simples de ser utilizado, o que o tornou um dos algoritmos mais utilizados. O algoritmo funciona criando uma “floresta” de arvores de decisões e por fim realiza uma combinação entre

elas, com o intuito de obter maior precisão. Cada árvore de decisão possui ramificações, nós e folhas, por onde serão feitas análises lógicas condicionais (se então). Vale notar que há uma diferença entre árvores de decisão e florestas aleatórias, enquanto a árvore de decisão, cria regras e nodos conforme aprende e recebe informações, a *Randon Forest*, faz isso aleatoriamente (DONGES, 2018).

Como vantagem do uso desse modelo, é que ele pode ser utilizado tanto para regressão quanto para classificação, ele é um algoritmo de fácil compreensão e é simples de identificar o grau de importância atribuído para cada entrada, além de ter a possibilidade de trabalhar com dados sem a necessidade de um pré-processamento. Contudo, caso ocorra de haver um número muito elevado de árvores no algoritmo, isso pode aumentar consideravelmente o tempo de processamento o tornando ruim para previsão de necessidade imediata (DONGES, 2018).

2.3. MODELO DE ASSOCIAÇÃO

2.3.1 Algoritmo A Priori

Entre as regras de associação, o algoritmo a priori é o que mais se destaca, principalmente por ser capaz de trabalhar com um grande número de atributos, o que resulta em uma vasta opção de alternativas combinatórias, e realiza essa tarefa com um baixo tempo de processamento (Agrawal& Srikant, 1994), foi inclusive, apontado como algoritmo mais promissor de geração de regras pela *International Conference on Data Mining* (ICDM) (Wu *et al*., 2007).

O algoritmo a priori, funciona encontrando conjunto de padrões, os conjuntos que acorrem com maior frequência são mantidos, e os com menor frequência são eliminados, avaliando as associações e retornando associações com base em critérios de confiança e suporte (ROMÃO, 2002).

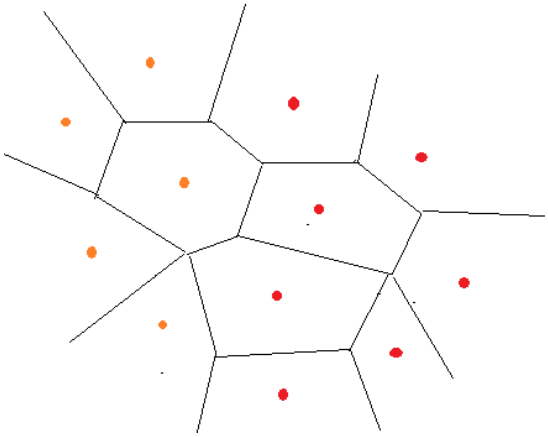
2.4. MODELO DE AGRUPAMENTO

2.4.1. K-Means

O K-means é um algoritmo de aprendizado não supervisionado, ele é um método de agrupamento (ou *clustering*), cujo divide os objetos de dados em grupos (k-grupos) onde cada uma das observações pertence ao grupo mais próximo da

média. Ao criar essas partições o K-means gera um diagrama de voronoi, cujo pontos centrais seriam como “centros de gravidade” ou centroides e as divisões seriam cada grupo, ilustrado na figura 07. (TAKAHASHI& BEDREGAL, 2005).

Figura 07 – Diagrama “*K-means”*



Fonte: Própria (2019)

Esse algoritmo pode ser utilizado em formar distintas, como para agrupar produtos com base na velocidade de venda deles ou do custo, ou então para agrupar cliente com base em métricas selecionadas. Uma das suas principais desvantagens é, cujo é um obstáculo comum em algoritmos com esse objetivo, é a dificuldade do algoritmo de quando encontra o mínimo local, encontrar o mínimo global, que seria o ideal. Outra dificuldade seria pela necessidade de identificar o que significa cada *cluster*, pois o algoritmo não demonstra, o que cabe a interpretação do analista (SAMPAIO, 2018).

2.4.2 DBScan

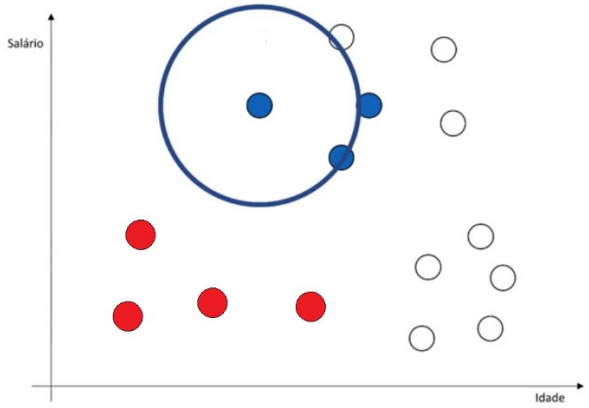
A Clusterização Espacial Baseada em Densidade de Aplicações com Ruído ou DBScan, é um método de agrupamento sem a necessidade de parametrização proposto por ESTER et al (1995).

O algoritmo DBScan trabalha de uma forma bem parecida como *K-means*, agrupando pontos similares no mesmo espaço, porém ele tem como vantagem, não ter a necessidade de o analista informar o número de *cluster* através de algum tipo de métrica, o próprio algoritmo já é capaz de realizar isso, além de ter uma velocidade de processamento maior se comparado como *K-means*. Como desvantagem do algoritmo, entra o fato dele não trabalhar bem quando lida com uma variedade muito

grande de densidades, caso essas densidades sejam umas muito próximas das outras (LUTINS, 2017).

O método de funcionamento do DBScan é relativamente simples, assim como ilustrado na figura 08, ele seleciona um ponto aleatório, traça uma circunferência ao redor desse ponto, de raio definido e os pontos q estão inclusos dentro dessa circunferência serão pertencentes a um agrupamento, o mesmo é realizado com cada ponto desse cluster, incluindo todos os pontos que estiverem dentro da circunferência, quando os pontos acabarem o algoritmo busca outro ponto aleatório e inicia o processo novamente (MAXWELL, 2016)

Figura 08 – "*DBScan*"



Fonte: Própria (2019)

3. MÉTODOS E TÉCNICAS DE PESQUISA

A natureza dessa pesquisa se enquadra como quanti-qualitativa. Para Creswell e Plano Clark (2011) esse método de pesquisa é a união das duas técnicas qualitativa e quantitativa em um único tipo de pesquisa. Pelo fato de o estudo na área de engenharia de produção possuir uma grande variedade de elementos e por explorar questões pouco estruturadas, a pesquisa quali-quantitativa se torna adequada (ENSSLIN e VIANNA, 2008).

Quanto ao objetivo da pesquisa, ela é de caráter explicativo “tem por objetivo explicar a razão das coisas” (PIOVESAN e TEMPORINI, 1995). Esse tipo de pesquisa propõe explicar de forma fundamental fenômenos (DUARTE, 2017).

Referente ao método utilizado nessa pesquisa, ela se enquadra como simulação e modelagem. A simulação é um método pouco utilizado nas ciências sociais, se comparado ao estudo de caso, ele necessita de um ambiente virtual. Essa metodologia geralmente utilizada para testar um modelo ou então realizar expectativas futuras para determinada situação, “a simulação se presta tanto no contexto de descoberta quanto no contexto da prova “ (VICENTE, 2004).

3.1. As etapas para realizar o trabalho

Para realizar as etapas do trabalho descritos acima, será utilizado a linguagem de programação Python em ambiente IDE Anaconda, com bibliotecas específicas de análise de dados.

4. DESENVOLVIMENTO DO TRABALHO

4.1 A EMPRESA

A empresa B, é uma empresa brasileira, de grande porte, fundada em 1945 e possui sedes no Brasil, Uruguai e Argentina. É uma empresa securitária que possui diversos serviços como seguro de vida, seguro de residência, seguro para animais, seguro odontológico, entre outros.

Apesar da diversidade de serviços ofertados, a maior parcela de faturamento da empresa é resultado do seguro de automóveis. Esse cenário ocorre por diversos fatores internos e externos, mas principalmente por ser um dos primeiros serviços que a companhia ofertou, portanto está enraizado e amadurecido dentro da empresa. Dessa forma, ela apresenta excelência na prestação desse serviço, visto que colhe resultados lucrativos até os dias de hoje.

Vale ressaltar também, a vantagem histórica brasileira para serviços automobilísticos. Esse serviço é impulsionado pela predominância rodoviária que o sistema logístico brasileiro oferece, ocasionando uma alta de demanda de seguros de automóveis. Vale ressaltar que de acordo com o Denatran, o Brasil tem uma média de um carro para quatro habitantes.

Com o objetivo de abordar o assunto ressarcimento no ramo de seguros, cabe abordar os principais termos técnicos da área.

Um dos principais termos utilizados no meio de seguros é a palavra sinistro. Um sinistro ocorre quando um bem do segurado sofre um prejuízo material, como por exemplo, quando um automóvel que está segurado pela companhia sofre um acidente de trânsito.

Outro termo também utilizado é o termo apólice de seguro. A apólice de seguro, é justamente o seguro de vida, ou de automóvel. Seria o documento emitido por uma seguradora. Portanto caso um indivíduo possua uma apólice de seguro, significa que ele possui um bem material segurado.

Quando se há pretensão de se referir ao indivíduo que não possui seguro dentro da companhia, geralmente esse indivíduo é referido como “terceiro”. Ou seja, em um acidente de trânsito que há dois indivíduos envolvidos e um deles sendo o segurado, o outro será referido como “terceiro”.

Outro jargão importante para ser descrito é o termo “ressarcido”. O termo “ressarcido” ou “passível de ressarcimento” refere-se a um sinistro cujo é possível reaver o prejuízo causado pelo terceiro em um bem material segurado pela empresa.

4.2 SEGUROS DE AUTOMÓVEIS E SEUS PROCESSOS

Com o objetivo de abordar o assunto de ressarcimento de valores dentro do cenário de uma seguradora, torna-se necessário descrever todo o fluxo de ressarcimento, desde a ocorrência de um sinistro, até o momento que o terceiro realiza o devido pagamento a companhia. Dessa forma, ficará mais claro, para tratar do assunto em questão e mensurar os ganhos desse trabalho.

O processo inicia com a ocorrência de um sinistro, por exemplo, um acidente de trânsito com uma colisão na traseira no veículo do segurado. Tratando de forma legal, acidentes de trânsito com colisão traseira, geralmente a responsabilidade não recai sobre o veículo que teve a traseira danificada e sim o outro. Dessa forma, como a responsabilidade do acidente não é do segurado, cabe ao terceiro arcar com os prejuízos do acidente. Portanto, esse tipo sinistro é direcionado para o setor ressarcimento.

Em resumo, o segurado aciona o seguro e a companhia cobre os gastos do acidente de imediato, independentemente de a responsabilidade ser ou não do segurado.

Esse sinistro, é encaminhado para análise da companhia e nessa etapa é avaliado se o sinistro é cabível de ressarcimento ou não. Essa análise é feita por analistas, com pouca utilização de rotinas automatizadas. São utilizadas informações como: descrição do segurado sobre o acidente por escrito no sistema, descrição do acidente por contato telefônico, análise da dinâmica acidente e conhecimento tácito do analista.

Caso o sinistro seja julgado como passível de ressarcimento, esse sinistro é encaminhado para o setor do ressarcimento e se inicia o processo de cobrança do terceiro. Caso o terceiro decida entrar em um acordo amigável, é gerado um acordo e o terceiro realiza o pagamento do débito para a companhia. Contudo, caso o terceiro negue o acordo amigável, é iniciado um processo judicial, e como fim do processo, o terceiro irá pagar ou não esse débito, de acordo com a decisão judicial.

O estudo desse trabalho é realizado nas etapas finais do processo, quando é iniciado o processo judicial, pois são coletados diversos tipos de dados que são cruciais para esse estudo, como dados do terceiro, região do processo judicial, dados dos advogados envolvidos, entre outros. Nessa etapa do processo, torna-se importante encontrar quais processos terão maior probabilidade de serem ressarcidos ou não, para reduzir ou evitar gastos desnecessários da companhia. Logo é a partir dessa necessidade e utilizando os dados citados que será aplicado o *machine learning*.

4.3 WEKA E PYTHON

Para desenvolver os modelos de *machine learning*, como se tem o intuito de aumentar o aprendizado, foi utilizado não só o software WEKA, mas também a linguagem Python no IDE Jupyter Notebook.

De acordo com Markov e Russel (2006), o WEKA é software que contém uma coleção de algoritmos de aprendizado de máquina. Com essa ferramenta é possível não só preparar os dados e aplicar modelos de regressão, agrupamento e regras de associação, mas também é possível visualizar os resultados obtidos.

Por se tratar de um software de código aberto, eficaz para processar *big data* e *deep learning* e possuir cursos gratuitos, tem sido amplamente utilizado pela comunidade de *data science*.

Para desenvolver os modelos, também foi utilizada a linguagem *Python*. Para Grus (2019), a linguagem *Python* é uma das melhores escolhas para lidar com tratamento e modelagem de dados. É uma linguagem gratuita, de rápido e simples aprendizagem e principalmente, possui diversas bibliotecas que auxiliam no desenvolvimento.

Ainda para Grus (2019), ainda que haja outras linguagens consideradas mais robustas, seguras ou estáveis, ao se trabalhar com dados com o objetivo de assertividade e velocidade de entrega sem perca da qualidade, a escolha acaba por ser a linguagem de programação *Python*.

4.4 APLICAÇÃO DO “*MACHINE LEARNING”*

Com o intuito de comparar métodos e verificar qual trará um melhor resultado para o problema proposto. Foi feito a modelagem com três métodos, sendo eles: Rede neural, árvores de decisão e regressão linear.

Nesse capítulo serão descritas as etapas do código desenvolvidos de forma resumida. Ou seja, não será comentado e explicado cada linha do código, e sim o intuito por trás de cada bloco de lógica de programação desenvolvido.

4.4.1 Rede Neural

Para o método de redes neurais, foi utilizado a linguagem *Python* no IDE *Jupyter Notebook*. Para realizar o desenvolvimento, como passo inicial, é necessário tratar os dados da base de dados. Inicialmente, foi necessário trocar valores ou “*flags*” como “Sim” e “Não” para valores numéricos, com “sim” igual a 1 e “não” igual a 0. (conforme figura 09).

Figura 09 – Tratamento dos dados com valores categóricos

```
1 my_data.head(3)
```

	Marca	Ramo	UF	Foro	Reu	Valor da Causa	Escritório	tipo_distribuicao	penhorado	suspenso	acordo	sentenca	cumprimento_de_sen
0	Porto	Dano Elétrico	SC	CENTRAL	CELESCDISTRIBUIÇÃOS.A.	3333.48	Sperotto Advogados Associados	SORTEIO	0	0	0	1	
1	Porto	Dano Elétrico	PR	CENTRAL	COPELDISTRIBUIÇÃOS/A-PARANÁ	8803.97	Sperotto Advogados Associados	SORTEIO	0	0	0	0	
2	Porto	Dano Elétrico	PR	FAZENDA PÚBLICA	COPEL	6118.41	Van Cleef	SORTEIO	0	0	0	0	

Fonte: do autor (2021)

Logo após, com um objetivo semelhante, é necessário utilizar uma técnica chamada de “*onehotencode*” para transformar campos que estão em texto em campos numéricos. Dessa forma o algoritmo consegue interpretar vetores de tipo *string*, como por exemplo um vetor de cidade que está presente na base de dados (figura 10).

Figura 10 – Aplicação do “*OnehotEncode*”

```
In [59]:
#ONEHOTENCODE

# seleciona para X apenas as colunas q sao objecto, ou seja as que vao ser transformados pelo hotenc
x_ohc_marca = X[['Marca']]
x_ohc_escritorio = X[['Escritório']]
x_ohc_ramo = X[['Ramo']]

x_ohc_foro = X[['Foro']]
x_ohc_uf = X[['UF']]
x_ohc_reu = X[['Reu']]
x_ohc_distribuicao = X[['tipo_distribuicao']]

#faz o onehotencod e ja passa para np array

#Reu

#faz uma lista com os valores unicos da coluna
x_ohc_reu = my_data['Reu'].values.tolist()
x_ohc_reu_unico = list(set(x_ohc_reu))


#cria um dicionario para guardar o valor numerico e oq ele representa
reu_mapa = {}
for x in range(len(x_ohc_reu_unico)):
    reu_mapa[x_ohc_reu_unico[x]] = x


#substitui a lista pelo número
for x in range(len(x_ohc_reu)):
    x_ohc_reu[x] = reu_mapa[x_ohc_reu[x]]

x_ohc_reu = to_categorical(x_ohc_reu)
```

Fonte: Própria (2020)

Portanto, cada vetor que estava como tipo texto, como por exemplo, o vetor de cidade, prestador ou de ramo, foi transformado em campos numéricos.

Agora, com a base de dados já tratada, foi utilizado a biblioteca “Keras”, para a aplicação do método de redes neurais (figura 11).

Figura 11 – Modelo Rede Neural

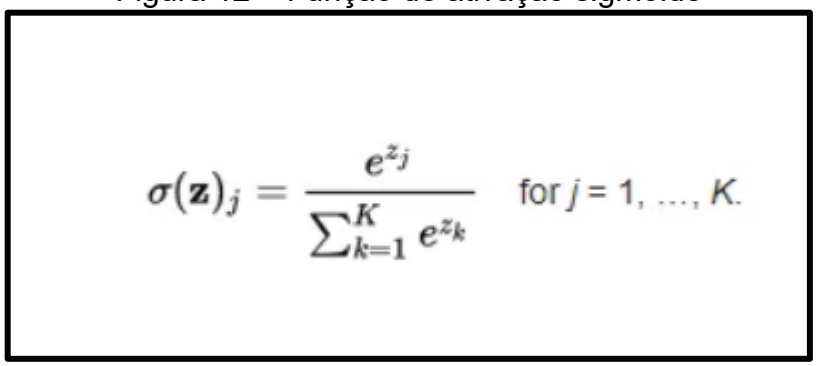
```
model = Sequential([
    Dense(units=152, input_shape=(303,), activation='relu'),
    Dense(units=76, activation='relu'),
    Dense(units=38, activation='relu'),
    Dense(units=2, activation='softmax')
])


model.compile(optimizer=Adam(learning_rate=0.01), loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

```
model.fit(x=X, y=y, validation_split=0.2, batch_size=10, epochs=100, shuffle=True, verbose=2)
```

Fonte: do autor (2021)

Para a base de dados em questão, foi utilizado três camadas na rede neural, a primeira com 152 neurônios, a segunda com 76 e a terceira com 38 neurônios. Com uma taxa de aprendizagem de apenas 0,01, para obter um melhor resultado e função de ativação linear. Já para a última camada foi utilizado a função de ativação sigmoide ou *softmax*. Que pode ser definida como na figura 12.

Figura 12 – Função de ativação sigmoide



Fonte: Própria (2020)

Ao executar o modelo com 1000 iterações, o resultado encontrado foi de cerca de 78% de acerto.

4.4.2 Regressão Logística

Para o método de regressão logística, foi utilizado o software WEKA. Como já citado, ele além de possuir os algoritmos para montar os modelos, ele também dispõe de ferramentas para tratamento de dados.

Como o algoritmo de regressão logística aceita rótulos nominais, na etapa de pré-processamento, foi utilizado o filtro de "*Numerictonominal*", para realizar essa transformação nos dados.

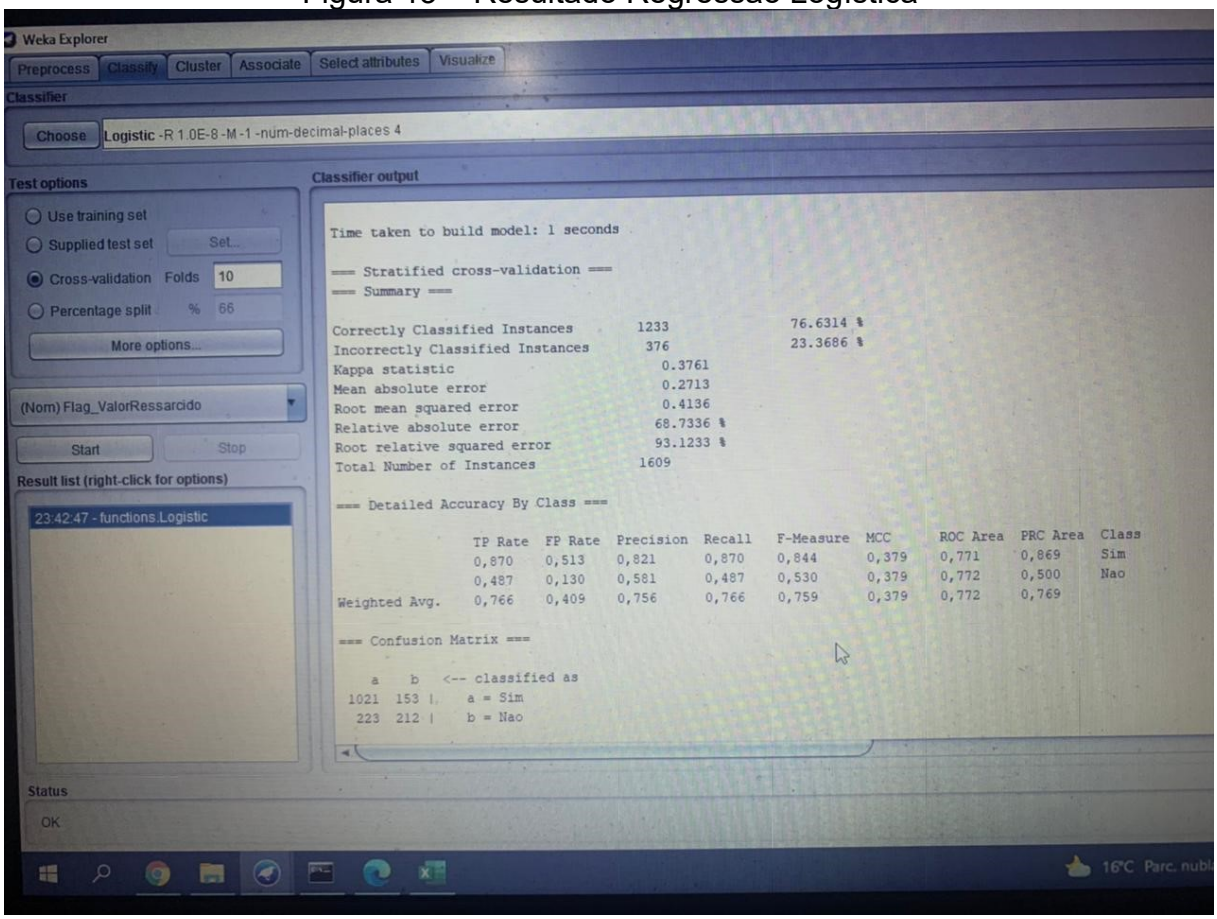
Com o intuito de normalizar os valores numéricos, também foi aplicado o filtro "*Normalizer*", assim evitando que valores muito divergentes, causasse um impacto na análise do algoritmo

Para que o algoritmo trate o vetor de data corretamente, é necessário transformar ele em numérico, portanto, também foi utilizado o filtro "*Datetonumeric*".

Com o mesmo objetivo que o "*onehotencode*" foi utilizado no algoritmo de redes neuras, no WEKA, está disponível o filtro "*StringtoWordVector"*, que faz justamente, a transformação de palavras em vetores que possibilitam a interpretação do algoritmo, logo esse filtro também foi aplicado a base de dados.

Após a etapa de pré-processamento ser concluída, foi aplicado o algoritmo de regressão linear, conforme a figura 13.

Figura 13 – Resultado Regressão Logística



Fonte: Própria (2020)

A regressão logística apresentou um acerto de 76,63% na classificação das instâncias, possuindo um grau de assertividade médio.

4.4.3 Árvores de Decisão

Para aplicar o algoritmo de árvore de decisão ou "*Randomforest*", também foi utilizado o software WEKA.

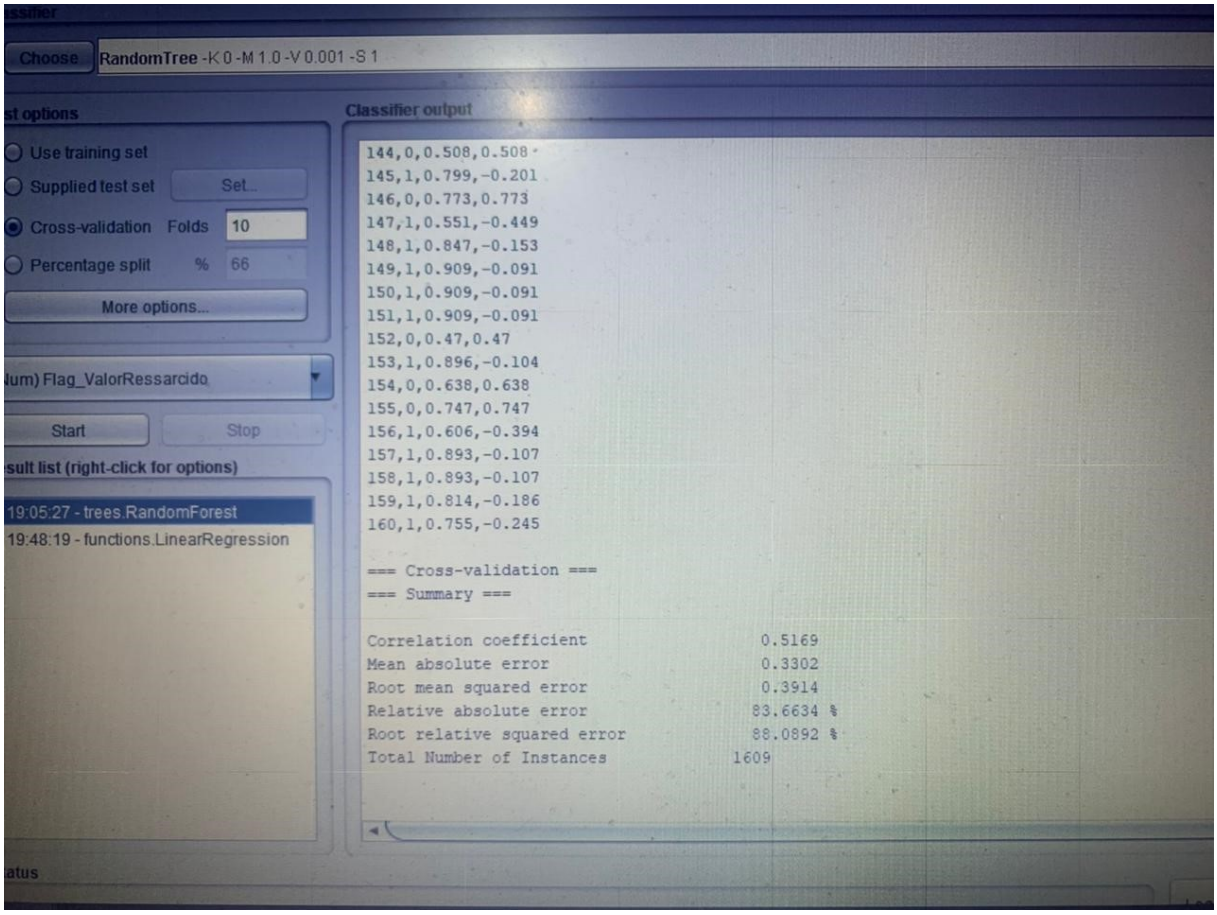
Na etapa de pré-processamento que o software oferece, como já citado, foi necessário transformar o vetor de valor, que estavam com os decimais com vírgula, para ponto, devido ao software aceitar apenas o modelo americano. E no mesmo vetor valor, também foi aplicado o filtro "*Normalizer*", para evitar erros na análise do algoritmo.

E assim como foi feito para o algoritmo de regressão logística, também foi utilizado "*Datetonumeric*", para tratar o vetor de data e o filtro de "*StringtoWordVector*"

para tratar vetor com nome de cidade ou de escritório de terceiros, como feito no algoritmo anterior.

Com a etapa de pré-processamento finalizada, foi aplicado o algoritmo de árvore de decisão, conforme figura 14.

Figura 14 – Resultado “*Randomforest”*



Fonte: do autor (2021)

O Random Forest apresentou um grau de correlação de aproximadamente 51,69%, ficando com um grau de assertividade baixo.

5. RESULTADOS E DISCUSSÕES

Para ser possível se obter uma boa previsão de quais processos possuem maior ou menor probabilidade de ser ressarcido, foram aplicados 3 algoritmos diferentes na mesma base de dados, com objetivo de comparação de entre si e para definir qual possui melhor qualificação para a base de dados. Os três algoritmos obtiverem taxas de acerto consideráveis e serão expostos para a companhia com o objetivo de contribuir com os resultados.

Com esse trabalho era esperado criar um modelo de previsão que fosse capaz de prever com alta taxa de acerto se um determinado processo judicial, iria ou não ter como resultado o valor ressarcido para a companhia.

Vale ressaltar, que a pesquisa auxiliou também a companhia a investir mais recursos para esse tipo de projeto, possibilitando novas oportunidades para aplicação de aprendizado de máquina em outros tópicos ou em bases de dados que a empresa já possui.

Nesse mesmo sentido, também surgiu uma maior motivação por parte da equipe em não só aumentar coleta de dados para esse tipo de estudo, mas também em manter os dados atualizados e certificar que estejam corretos.

Ao longo da pesquisa, foi possível notar que por maior que fosse a quantidade de dados e variáveis que o modelo possuía, um dos maiores desafios foi superar a jurisprudência que possui um forte papel nesse tipo de processo, pois impacta na decisão do júri de uma forma que o modelo de previsão desenvolvido tem maior dificuldade em lidar.

Para o método de redes neurais, foi obtido um resultado de 78% de acerto. Para árvore de decisões, se obteve um grau de correlação de 51,69% e para a regressão logística também se obteve um resultado próximo das redes neurais, com um grau de assertividade de 76,63%.

Como resultado, esperava-se um acerto menor do resultado do aprendizado de máquina, pelo fato de a base de dados não possuir um número tão alto de amostras e pelos processos não seguirem um padrão claro de serem passíveis de ressarcimento ou não.

Se por um lado, a base de dados possuía informações que auxiliam na resposta do julgamento de forma assertiva. Vale levar em consideração que, algumas "*features*" da base de dados, possuíam maior impacto no resultado final, apenas levando em consideração a localidade do processo judicial, já era obter uma previsão com um resultado considerável.

Nota-se que para a base de dados em questão, o algoritmo possuía não só informações da localidade, mas outras informações como, o valor do processo, que pode ocasionar em maior ou menor esforço do escritório parceiro no processo. Mas possuía também qual o réu envolvido, qual prestador iria conduzir o processo judicialmente e informações de em qual situação o processo se encontra, como por exemplo, se já ocorreu penhora dos bens ou se o terceiro já foi citado ou não.

Porém, por outro lado, vale notar que um dos principais desafios para se obter uma alta taxa de acerto, seria que nesse tipo de processo judicial, a questão da jurisprudência acaba por ter um impacto definitivo no resultado do julgamento, como já citado anteriormente. Portanto, para o mesmo processo, o resultado pode variar de juiz para juiz. Ainda que esse requisito tenha sido relativamente incluso para a decisão dos algoritmos, já que na base de dados, discriminava qual era o juiz para aquele processo. A decisão humana, acaba sendo um fator difícil de uma máquina prever, principalmente porque está passível a diversos fatores externos.

Outro ponto que ajudaria na obtenção de melhores resultados, seria uma base de dados com um maior número de dados, como por exemplo a informação da pontuação do terceiro no Serasa, ou informações mais especificas do terceiro, porém, a empresa segue uma política rígida de privacidade de dados e nem todas as informações estão disponíveis para a utilização.

Outro fato importante, é que a base de dados não era robusta em quantidade de linhas, quanto o desejado, continha cerca de 1500 linhas, o que limita o aprendizado dos algoritmos. Porém, ainda assim, foi possível obter bons resultados.

Outro ponto positivo, é que esse tipo de coleta de informações se iniciou recentemente no setor ressarcimento da companhia. E a pesquisa motivou não só a continuar com a coleta de dados, como a buscar por mais informações, para aplicar melhorias esse modelo novamente e para outras frentes de pesquisa. Portanto, ainda que os resultados dos algoritmos aplicados não tenham sido com uma taxa de acerto

acima de 90%, o estudo trouxe outros benefícios, tanto para a companhia quanto para aprendizado para o pesquisador.

6. CONCLUSÃO

O desenvolvimento do presente estudo possibilitou uma análise, aplicação e comparação de resultados de três tipos diferentes algoritmos, para desenvolver um modelo preditivo e definir se um processo judicial é cabível de ressarcimento ou não. Além disso, também permitiu um aumento da coleta e organização dos dados dentro da companhia e motivou as partes envolvidas a realizar outras pesquisas com o mesmo intuito desse trabalho e a aumentar coleta de dados da companhia.

Nesse sentido, com os objetivos iniciais da pesquisa cumpridos, sendo alcançado uma capacidade de predição de nível mediano, cerca de 77% de acerto. E de construir um modelo de aprendizado de máquina capaz de predizer o resultado de um processo judicial, identificou-se também a importância da utilização de ferramental matemático dentro do ambiente organizacional e seus processos.

Por fim, essa pesquisa abre espaço para as descobertas do estudo e a possibilidade de investigações futuras em outras bases de dados da companhia. Em pesquisas futuras, pretendo desenvolver uma interface para que o usuário possa inserir os dados do processo e obter uma resposta não só do resultado, mas a probabilidade de acerto do algoritmo.

7. REFERÊNCIAS

PRATI, Ronaldo Cristiano. **Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos**. 2006. Tese de Doutorado. Universidade de São Paulo.

GARRAIO, Joana Rita Barradas. **Modelação da taxa de anulação no seguro automóvel**. 2015. Tese de Doutorado.

SANTOS, Ana Cíntia Brandão dos; FARIAS, Gilmar Alves de. Reudes neurais: um conceito para matemática industrial. 2019.

CARVALHO, Maristela Rodrigues de. A contribuição da mineração de dados na recuperação da informação e suas relações com a biblioteconomia. 2019.

FREITAS, Igor Wescley Silva de et al. Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados. 2019.

BERNARDI, Élder Francisco Fontana et al. Uma arquitetura para suporte à mineração de dados paralela e distribuída em ambientes de computação de alto desempenho. 2010.

HEARST, Marti A.. et al. Support vector machines. IEEE Intelligent Systems and their applications, v. 13, n. 4, p. 18-28, 1998.

GRUS, Joel. **Data Science do zero: Primeiras regras com o Python**. Alta Books, 2019. https://books.google.com.br/books?hl=pt-BR&lr=lang_pt&id=x4-wDwAAQBAJ&oi=fnd&pg=PT5&dq=python&ots=Lv_fm8ss85&sig=pRFu0HNqabgYca8S_1ekXdwxNik#v=onepage&q&f=false

REZENDE, Solange Oliveira. **Sistemas inteligentes: fundamentos e aplicações**. Editora Manole Ltda, 2003.

EHLERS, Ricardo Sandes. Introdução à Inferência bayesiana. **URL: http://www. leg. ufpr. br/% 7Epaulojus/CE227/ce227. pdf**, 2003.

Figueira, Cleonis Viater. "Modelos de regressão logística." (2006).

Markov, Zdravko, and Ingrid Russell. "An introduction to the WEKA data mining system." ACM SIGCSE Bulletin 38.3 (2006): 367-368.

LORENA,A.C.;CARVALHO,A.C.P.L.F. Uma Introdução às Support Vector Machines.Revista de Informática Teórica e Aplicada,v.14, n.2, p.43-67, 2007.

BREIMAN, Leo et al. Classification and regression trees. Wadsworth Int. **Group**, v. 37, n. 15, p. 237-251, 1984.

FERNEDA, Edberto. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, v. 35, p. 25-30, 2006.

Turban E., Sharda R., Aronson J. E, King D. “Business Intelligence: Um enfoque gerencial para Inteligência do negócio”. Porto Alegre: Bookman, 2009, 1 ed., 250p.

SILVEIRA, Juliano Gomes. RSAPP, Um algoritmo baseado em *Rogh Sets* para o auxílio ao processo de descoberta de conhecimento em banco de dados. **2013**, Universidade Católica do Rio Grande do Sul, 2013. http://repositorio.pucrs.br/dspace/handle/10923/5544?mode=full#preview

Chakrabarti, Soumen, et al. "Data mining curriculum: A proposal (Version 1.0)." *Intensive Working Group of ACM SIGKDD Curriculum Committee* 140 (2006): 1-10.

SANTOS, Lucas Costa Oliveira. Aplicaçao do Processo de KDD a um Ambiente Industrial. 2007.

STOROPOLI, José Eduardo et al. O uso do Knowledge Discovery in Database (KDD) de informações patentárias sobre ensino a distância: contribuições para instituições de ensino superior. 2016.

KINAS, Paul Gerhard; ANDRADE, Humber Agrelli. **Introdução à Análise Bayesiana (Com R)**. 1. ed. [*S. l.*]: Consultor Editorial, 2017. 295 p.

SOUZA, Isaac Jales Costa. **Estimação Bayesiana no Modelo Potência Normal Bimodal Assimétrico**. 2016. Tese (Pós-Graduação em Matemática Aplicada e Estatística) - Universidade Federal do Rio Grande do Norte, Natal, 2016. https://repositorio.ufrn.br/jspui/bitstream/123456789/21722/1/IsaacJalesCostaSouza_DISSERT.pdf.

GAMA, J. Árvores de decisão. 2000. Disponível em: <http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html>. Acesso em: 31 ago. 2019

WITTEN, I.H.; FRANK, E. Data mining: pratical machine learning tools and techniques with Java implementations. San Francisco, California: Morgan Kaufmann, 2000

PONNIAH, P. Data Warehousing Fundamentals: A Comprehensive Guide for IT

Professionals. John Wiley and Sons, Inc, 2001

COLARES, Peterson Fernandes et al. Processo de indução e ranqueamento de árvores de decisão sobre modelos OLAP. 2010.

INGARGIOLA, Giorgio. Building Classification Models: ID3 and C4.5. Disponível por WWW em: http://www.cis.temple.edu/~ingargio/cis587/ readings/id3-c45.html, 1996.

BRADZIL, P. B. Construção de modelos de decisão a partir de dados. 1999. Disponível em: <http://www.nacc.up.pt/~pbrazdil/ Ensino/ML/ModDecis.html>. Acesso em: 31 ago. 2019.

MARTINS, Marco Antônio dos Santos; METTE, Frederike; MACEDO, Guilherme Ribeiro. A Utilização de Redes Neurais Artificias para a Estimação dos Preços da Petrobrás PN na Bovespa. ConTexto, Porto Alegre, v. 8, n.14, 2o semestre 2008.

TAFNER,Malcon Anderson. Redes Neurais Artificiais: Aprendizado e Plasticidade. Revista “Cérebro & Mente”, 2-5 mar/mai, 1998

FINOCCHIO, Noções de Redes Neurais Artificiais. Paraná, UTFPR, 2014.

CAMARGOS, M. A.; CAMARGOS, M. C. S.; ARAÚJO, E. A. A inadimplência em um programa de crédito de uma instituição financeira pública de Minas Gerais: uma análise utilizando regressão logística. **Revista de Gestão**, v. 19, n. 3, p. 467-486, 2012.

Smola, A. J., Barlett, P., Sch¨olkopf, B., and Schuurmans, D. (1999b).Introduction to Large Margin Classifiers, capítulo 1, paginas 1–28. I nSmola et al. (1999a).

Cristianini, N. and Shawe-Taylor, J. (2000).An Introduction to Support Vector Machinesand other kernel-based learning methods. Cambridge University Press.

Lorena, A. C. de Carvalho, A. C. P. L. F. (2003) Introdução às Máquinas de Vetores Suporte (Support Vector Machines). Relatório Técnico nº 192 do Instituto de Ciências Matemáticas e de Computação da USP

SOARES, Heliana Bezerras. Análise e classificação de imagens de lesões da pele por atributos de cor, forma e textura utilizando máquina de vetor de suporte. Orientador: Prof. Dr. Adrião Duarte Dória Neto. 2008. Tese (Pós-graduação em Engenharia Elétrica e Computação da) - Universidade Federal do Rio Grande do Norte, Natal, 2008. https://repositorio.ufrn.br/jspui/handle/123456789/15118.

KOTLER, Philip. Administração de marketing. 10aed. São Paulo:Prentice Hall, 2000. 764 p.

GARGANO, Michael L.; RAGGAD, Bel G., 1999. **Data mining - a powerful information creating tool.** OCLC Systems and Services, MCB University Press, v. 15, n. 2, págs. 81-90.

CAMILO, Cassio Oliveira; SILVA, João Carlos. Mineração de Dados: Conceitos, Tarefas, Métodose Ferramentas. [S. l.: s. n.], 2009. Disponível em: http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acesso em: 22 nov. 2019.

CAMILO, Cassio Oliveira et al. Uma Metodologia para Mineração de Regras de Associação Usando Ontologias para Integração de Dados Estruturados e Não-Estruturados. 2010.

Wu et al., 2007) X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg. (2007) "Top 10 algorithms in data mining". Knowledge and Information Systems, Vol. 14 Issue 1, pp. 1–37.

TAKAHASHI, Adriana; BEDREGAL, Benjamin René Callejas; LYRA, Aarão. Uma versão intervalar do método de segmentação de imagens utilizando o k-means. **Trends in Applied and Computational Mathematics**, v. 6, n. 2, p. 315-324, 2005.

Ester M., Kriegel H.-P., and Xu X. 1995. A Database Interface for Clustering in Large Spatial Databases, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, Montreal, Canada, 1995,AAAI Press, 1995.

ENSSLIN, L.; VIANNA, W. B. O desing na pesquisa quali-quantitavi em engenharia de produção - questões epistemológicas. **Revista Produção Online**, Santa Catarina, v. 8, n. 1, mar 2008

CRESWELL, J. W.; PLANO CLARK, V. L. **Designing and conducting mixed methods research**. Los Angeles: SAGE Publications, 2011.

DONGES, N. Towards Data Science. **Towards Data Science**, 2018. Disponivel em: <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>. Acesso em: 23 nov. 2019.

DUARTE, V. M. D. N. Brasil Escola. **Monografia Brasil Escola**, 2017. Disponivel em: <https://monografias.brasilescola.uol.com.br/regras-abnt/pesquisas-exploratoria-descritiva-explicativa.htm>. Acesso em: 21 Novembro 2019.

LUTINS, E. Medium. **Medium**, 2017. Disponivel em: <https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818>. Acesso em: 23 Novembro 2019.

MAXWELL. PUC - Rio. **Maxwell**, 2016. Disponivel em: <https://www.maxwell.vrac.puc-rio.br/24787/24787_6.PDF>. Acesso em: 23 Novembro 2019.

PETENATE, M. Escola EDTI. **Escola EDTI**, 2019. Disponivel em: <https://www.escolaedti.com.br/o-que-e-regressao-linear-entenda-aqui>. Acesso em: 23 Novembro 2019.

PIOVESAN, A.; TEMPORINI, E. R. Pesquisa exploratória: procedimento metodológico parao estudo de fatores humanos no campo da saúde pública. **Saúde Pública**, São Paulo, Maio 1995.

SAMPAIO, P. E. Medium. **Medium**, 2018. Disponivel em: <https://medium.com/@paulo_sampaio/entendendo-k-means-agrupando-dados-e-tirando-camisas-e90ae3157c17>. Acesso em: 23 Novembro 2019.

THOMAS, G. **Regressão Não Linear**. Universidade de São Paulo. [S.l.]. 2016.

VICENTE, P. O uso de simulação como metodologia de pesquisa em ciências sociais. **EBAPE**, Rio de Janeiro, Novembro 2004.

ZEVIANI, W. M.; JÚNIOR, P. J. R.; BONAT, W. H. **Modelos de regressão não linear**. Universidade Federal do Paraná. Campina Grande, p. 98. 2013.