UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA METALÚRGICA E DE MATERIAIS

CURSO DE GRADUAÇÃO EM ENGENHARIA METALÚRGICA

LUCAS GOMES MARTINS

APLICAÇÃO DE UM MODELO DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE CHURN: UM ESTUDO DE CASO

FORTALEZA

2021

LUCAS GOMES MARTINS

APLICAÇÃO DE UM MODELO DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE CHURN: UM ESTUDO DE CASO

Trabalho de conclusão de curso apresentado ao curso de Engenharia Metalúrgica do Departamento de Engenharia Metalúrgica e de Materiais da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia Metalúrgica.

Orientador: Prof. Dr. Ing. Jeferson Leandro Klug.

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M344a Martins, Lucas Gomes.
Aplicação de um modelo de aprendizagem de máquina para predição de churn : um estudo de caso / Lucas Gomes Martins. – 2021.
32 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Metalúrgica, Fortaleza, 2021.
Orientação: Prof. Dr. Jeferson Leandro Klug.

1. Floresta aleatória. 2. Aprendizado de máquina. 3. Modelo preditivo. I. Título.

CDD 669

LUCAS GOMES MARTINS

APLICAÇÃO DE UM MODELO DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE CHURN: UM ESTUDO DE CASO

Trabalho de conclusão de curso apresentado ao curso de Engenharia Metalúrgica do Departamento de Engenharia Metalúrgica e de Materiais da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia Metalúrgica.

Aprovada em: ___/___/_____.

BANCA EXAMINADORA

_____
Prof. Dr. Ing. Jeferson Leandro Klug (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Marcelo Ferreira Motta
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Jorge Luiz Cardoso
Universidade Federal do Ceará (UFC)

A Deus.

Aos meus pais, Lourdes e Ricardo.

AGRADECIMENTOS

Ao SENHOR, pois sem sua misericórdia nada seria.

Aos meus pais por todo amor e carinho durante minha vida.

Aos meus amigos de curso, em especial, Daniel Santos, Bruno Ribeiro, Saynarah Cruz, Matheus Souza e Cesanildo Sousa por terem sido meus companheiros e encorajadores durante toda a graduação.

Ao meu companheiro de labuta Magno, por toda ajuda.

Ao meu amigo Raimundo Balada por ter me ajudado em situações adversas.

Ao meu amigo Rafael Gomes por todos os conselhos e ensinamentos a mim repassados.

As minhas orientadoras enquanto fui bolsista, Nathália e Roberta, por toda paciência e companheirismo.

Aos gestores, Artur e Takano, que me deram a primeira oportunidade no mercado de trabalho.

Ao Prof. Dr. Ing. Jeferson Leandro Klug, pela excelente orientação.

Aos professores do Departamento de Engenharia Metalúrgica e de Materiais pelos ensinamentos e vivências repassadas, em especial para o professor Elineudo Pinho de Moura por ser um exemplo de docente a ser seguido.

“O insucesso é apenas uma oportunidade para recomeçar com mais inteligência.”
Henry Ford

RESUMO

O aprendizado de máquina é um algoritmo computacional de análise de dados que automatiza a construção de modelos analíticos. O campo de aplicação desses modelos é bastante vasto, sendo utilizado na engenharia, medicina, segurança pública, operações de crédito e diversas outras áreas. O presente trabalho utiliza o modelo de floresta aleatória para classificar clientes de uma empresa varejista que deixarão de utilizar os serviços prestados. A base de dados utilizada foi uma amostra do quantitativo total de clientes e é composta por todos os clientes que tiveram seus serviços cancelados nos últimos seis meses e clientes ativos escolhidos aleatoriamente. O conjunto de dados utilizado foi dividido em dois outros conjuntos, sendo um para treino e outro para teste do modelo preditivo. A escolha do modelo de floresta aleatória mostrou-se bastante assertiva uma vez que as estatísticas de performance e validação do algoritmo reforçam a força preditiva do modelo, apresentando uma acurácia de 98% e precisão de 97%.

Palavras-chave: Floresta aleatória. Aprendizado de máquina. Modelo preditivo.

ABSTRACT

Machine learning is a computational data analysis algorithm that automates the construction of analytical models. The field of application of these models is quite vast, being used in engineering, medicine, public security, credit operations and several other areas. The present work uses the random forest model to classify customers of a retail company that will no longer use the services provided. The database used was a sample of the total number of customers and is composed of all customers who had their services canceled in the last six months and active customers randomly chosen. The data set used was divided into two other sets, one for training and the other for testing the predictive model. The choice of the random forest model proved to be quite assertive since the algorithm's performance and validation statistics reinforce the predictive strength of the model, presenting an accuracy of 98% and a precision of 97%.

Keywords: Random Forest. Machine Learning. Predictive Model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CRM	*Client Relationship Managemenent*
ETC	Extração, Transformação e Carga
D	Conjunto de Dados
FN	Falso Negativo
FP	Falso Positivo
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machine*

LISTA DE SÍMBOLOS

% Porcentagem

® Marca Registrada

SUMÁRIO

1 INTRODUÇÃO

No mundo globalizado e com o avanço da tecnologia, milhares de dados e informações sobre processos industriais, padrões de comportamento, variações cambiais, segurança pública, dentre outros são gerados todos os dias. Assim, devido a limitação humana de absorver, organizar e gerar informação e/ou conhecimento da infinidade de dados gerados, foram desenvolvidos algoritmos computacionais que compilam, aprendem e geram informação com uma capacidade muito superior a humana. Diversos modelos já foram desenvolvidos como os modelos de regressão logística, árvores de decisão, redes neurais e florestas aleatórias. Na metalurgia esses modelos têm sido utilizados para prever falhas estruturais em equipamentos metálicos (DANTAS, 2015), medir o teor de ferro no concentrado da flotação do minério de ferro (GUEDES, 2020), melhorar a qualidade e rendimento de carvões vegetais (PEREIRA, 2019), predição de inclusões não metálicas em aços (MARTÍNEZ, 2019) e diversas outras aplicações. Para o caso particular desse trabalho, devido a disponibilidade dos dados, um modelo preditivo será usado para prever clientes que deixarão de consumir serviços de uma empresa de telecomunicações.

Com a entrada de novas companhias em vários nichos de mercado, é inevitável que ocorra o acirramento da concorrência entre elas. Desse modo, variados tipos de mercado, passam a ficar cada vez mais saturados e pressionados pelo aumento da competitividade (PIMENTEL, 2019). Como resultado, as empresas vêm notando que suas estratégias comerciais devem priorizar a manutenção dos clientes atuais, ao invés de atraírem novos (COUSSEMENT; POEL, 2009). Nessa perspectiva, existe um aumento da relevância dada às iniciativas de gerenciamento com o consumidor (em inglês, CRM – *Client Relationship Managemenent*) dentro das organizações. O CRM é uma abordagem de gerenciamento que visa desenvolver, aprimorar e criar os relacionamentos com clientes criteriosamente segmentados para maximizar a rentabilidade corporativa e o valor do cliente (A. PAYNE, 2005). Um dos maiores desafios enfrentados pelos CRM é a identificação de clientes propensos ao *churn* (i.e., cancelamento) de serviços e/ou produtos (HADDEN et al., 2007).

Além disso, é notório citar que existem estratégias que são utilizadas ao longo do ciclo de vida do cliente (GARCÍA et al., 2017):

- Aquisição: O momento em que o consumidor adquire um novo produto ou serviço, ou seja, inicia o seu relacionamento com a empresa;
- Fidelização: Esse é o momento em que a empresa identifica valor no cliente e de forma proativa tenta maximizar a extração de uma maior margem de

lucro pela sua permanência na base ou com a oferta de um produto novo da mesma empresa;

- Retenção: Nesse momento, o cliente aborda a empresa com o objetivo de realizar o *churn* e em contrapartida, a empresa de forma reativa tenta manter o cliente na base através de uma readequação do serviço e/ou produto a um custo menor para o cliente ou aplica descontos agressivos.

Com o objetivo de manter a base de clientes preservada, os analistas de relacionamento redirecionam sua atenção para os clientes com maior probabilidade de *churn*. As empresas estão orientando suas estratégias para o foco no cliente em detrimento do marketing de massa (BUREZ; VAN DE POEL, 2007).

As empresas e seus clientes estão em processo de evolução e isto pode levar a fragmentações naturais no relacionamento comercial. Logo, nestes casos a continuidade do ciclo de vida dos clientes não será natural, levando ao entendimento que nem todo *churn* é previsível (PIMENTEL, 2019). Desse modo, é necessário que exista um CRM e por consequência uma forma de gerenciar o *churn* do cliente. Portanto, é crucial classificar todos os tipos possíveis de *churn* (GARCÍA et al., 2017):

- *Churn* involuntário: Afeta clientes em que a empresa retira o serviço (por exemplo, inadimplência, fraude).
- *Churn* voluntário: São os clientes que por decisão particular decidem cancelar o serviço.

Diante do exposto, o modelo de aprendizagem de máquina terá como objetivo cumprir os requisitos imprescindíveis em um sistema de previsão de *churn*. Segundo Balle, Casas, & Catarineu (2011) os requisitos são os seguintes:

- Precisão: Na avaliação do classificador este deve possuir uma medida de *recall* elevado (pelo menos todos os *churners* são identificados) e precisão elevada (não existir muitos falsos positivos);
- Desempenho: A velocidade com que o modelo pode ser executado com novos dados é essencial para as decisões serem tomadas em tempo hábil;
- Escala: O modelo deve reagir de forma coesa com o aumento dos dados de alimentação;
- Flexibilidade: O modelo deve manter-se com bons índices de previsão com a alteração nos padrões dos clientes;

- Segmentação: Capacidade de serem retirados dados concretos sobre o perfil de clientes mais propensos a abandonarem o serviço;

2 OBJETIVOS

Este trabalho visa antecipar o *churn* de clientes através da aplicação de um modelo de aprendizagem de máquina utilizando variáveis de faturamento, cadastro e atendimento dos clientes de uma empresa de telecomunicações. É desejado identificar, com precisão, quais clientes deixarão a empresa, através de um modelo de classificação.

3 REFERENCIAL TEÓRICO

Nessa seção será abordado sobre como o processo ETC é utilizado para preparar uma base de dados bruta para a correta utilização em um modelo de aprendizagem de máquina e em como o modelo utiliza essa base de dados para fazer análises preditivas.

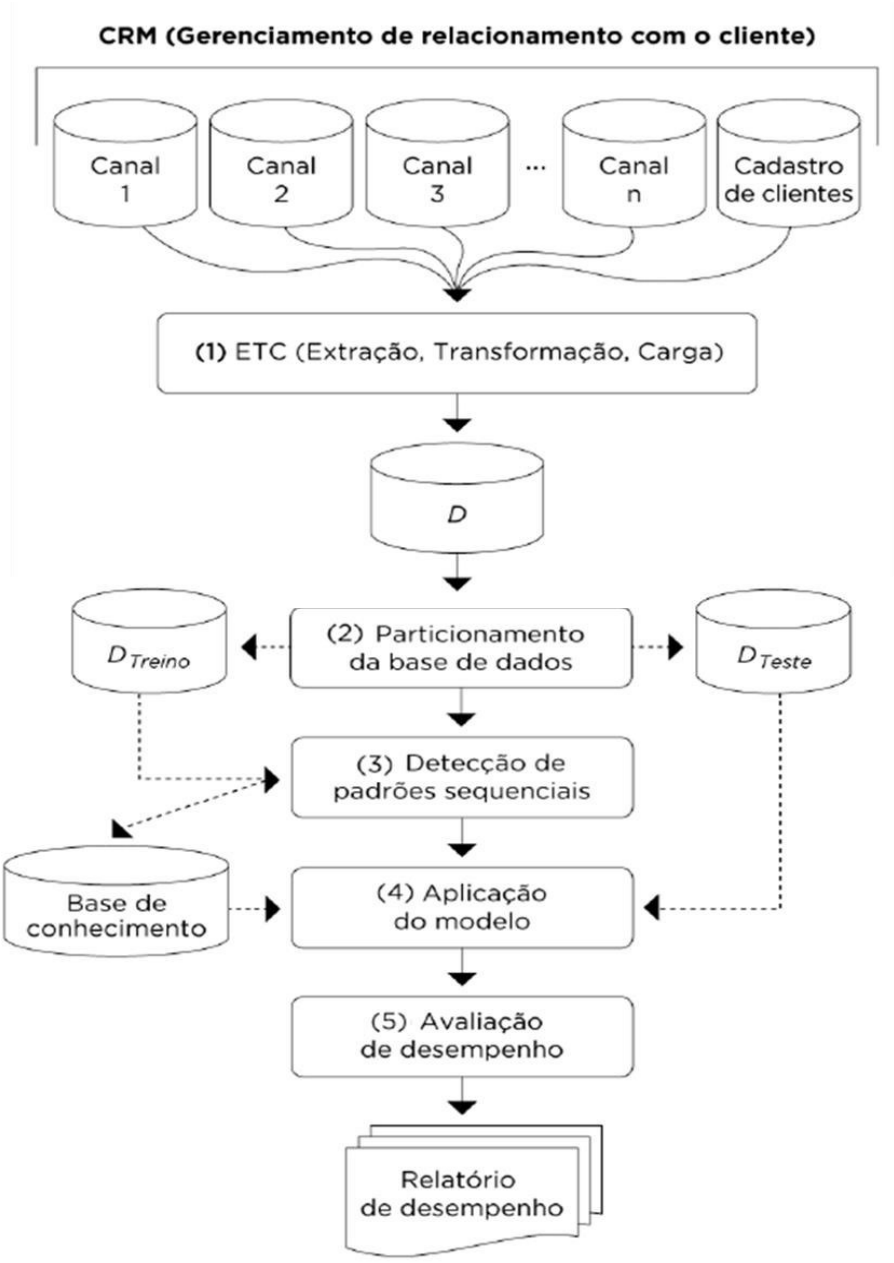
3.1 Extração, Transformação e Carga (ETC) de Dados

Processo responsável por realizar as operações de pré-processamento nos conjuntos de dados antes de sua utilização no modelo de aprendizagem de máquina. As operações contemplam a utilização de técnicas de seleção, como a eliminação manual de atributos, em que se observa como os atributos classificados como irrelevantes podem ser identificados e eliminados manualmente do conjunto de dados. Além disso, modificações para codificação dos tipos de variáveis, limpeza e integração dos dados também são utilizados.

A técnica de eliminação manual de atributos é a primeira técnica a ser aplicada no conjunto de dados, pois nem todas as variáveis e características do conjunto de dados original são necessários para utilização no modelo de aprendizagem de máquina (PIMENTEL, 2019). Quando o atributo não contribui de forma clara para a estimativa do valor alvo, ele pode ser considerado como irrelevante. Por exemplo, se temos um conjunto de dados com a idade, classe social, altura, peso, raça, sexo e biotipo de vários indivíduos e queremos utilizar um modelo para classificar estes em atletas ou não atletas, as variáveis classe social e signo podem ser descartas, uma vez que não possuem relevância para identificação de atletas.

A etapa de integração de dados é fundamental para a correta utilização dos dados, pois partindo de um CRM, os conjuntos de dados extraídos de diversos canais de relacionamento estão integrados. Desse modo, cada conjunto de dados pode estar representando diferentes atributos de um mesmo grupo de objetos. Assim, é de extrema importância, na integração de dados, identificar quais os objetos estão presentes nos conjuntos de dados a serem combinados. Por exemplo, dado dois conjuntos de dados, o primeiro conjunto possui o registro geral, idade e cor da população de uma cidade, já o segundo conjunto de dados possui o registro geral, sexo e religião da população da mesma cidade, a integração dos dois conjuntos de dados será feita através da variável de registro geral, que é comum aos dois grupamentos de dados. Algumas problemáticas podem gerar dificuldade na integração correta dos dados, como por exemplo, variáveis que são correspondentes podem possuir nomes diferentes em base de dados distintas. A figura 1 ilustra esse processo.

Figura 1 – Esquema do processo *ETC* e suas etapas



Fonte: PIMENTEL (2019, com adaptações).

Na etapa de limpeza de atributos, utiliza-se a validação da consistência das informações e a correção de erros existentes. É corriqueiro que alguns objetos associados aos conjuntos de dados estejam inconsistentes e incompletos. Identificar e remover dados inconsistentes é fundamental para a melhoria do resultado do modelo. A participação do especialista do domínio da aplicação é fundamental na limpeza de atributos (GOLDSCHMIDT; PASSOS, 2015). Por exemplo, um conjunto de dados com informações dos clientes de uma empresa que possui somente clientes no estado do Ceará, apresenta registros de clientes no

estado do Piauí. Para esse caso, a correção do estado dos clientes deve ser feita, alterando a informação que está registrada no conjunto de dados.

A técnica de codificação de atributos transforma os valores de determinados atributos do conjunto de dados. Nessa etapa, é crucial compreender que as variáveis devem ser codificadas para melhor atender as necessidades dos algoritmos utilizados nas etapas pós pré-processamento (GOLDSCHMIDT; PASSOS, 2015). Por exemplo, um conjunto de dados que representa o sexo de um grupo de pessoas como M (masculino) e F (feminino) será utilizado por um modelo preditivo que somente aceita variáveis numéricas. Nessa situação, a representação do sexo das pessoas deve ser substituída por um número.

Após todas as etapas aplicadas, as bases são compiladas em um novo conjunto de dados que, para o caso particular, foi chamado de *D*.

3.2 Aprendizado de Máquina

Aprendizado de máquina é um ramo da inteligência artificial que tem por objetivo o desenvolvimento de técnicas computacionais sobre o aprendizado e confecção de sistemas capazes de adquirir conhecimento de forma automatizada (MONARD; BARANAUSKAS, 2003).

Ainda que o aprendizado de máquina seja uma excelente ferramenta para conquista de conhecimento, não existe algoritmo que descreva, com qualidade, todos os problemas. Assim, é necessário estudar a problemática e aplicar o modelo que melhor descreva o problema.

Os modelos de aprendizado de máquina podem ser feitos de forma supervisionada ou não supervisionada. Os algoritmos de forma supervisionada aprendem através de instâncias externas de forma a produzirem hipóteses, que posteriormente fazem estimativas e previsões sobre instâncias futuras, em que cada exemplo é descrito por um vetor de atributos (KOTSIANTIS, 2007). O objetivo do modelo é construir um classificador que possa determinar de forma assertiva novos exemplos ainda não conhecidos. Já os modelos de forma não supervisionada necessitam entender os padrões, relações, categorias ou regularidades nos dados que lhe vão sendo apresentados e codificá-los nas saídas (ROJAS, 1996). O indutor analisa os exemplos fornecidos e determina se eles podem ser agrupados (CHEESEMAN & STUTZ, 1990).

Os valores de saída de um conjunto de dados utilizados em um modelo de aprendizagem supervisionada podem ser contínuos ou discretos. A distinção entre os valores de saída é importante para saber que tipo de problema o modelo irá resolver. Os principais

problemas são os de classificação e regressão (RAUTIO, 2019). Nos problemas de classificação as saídas são discretas, já para os problemas de regressão as saídas são valores numéricos e contínuos.

3.2.1 Modelo utilizado

O modelo de aprendizagem de máquina utilizado é o de floresta aleatória (em inglês, *random forest*). A escolha foi feita em virtude do modelo performar muito bem na classificação de *churn*, tal constatação pode ser observada na tabela 1, em que mais de 50% dos trabalhos utilizam o modelo de floresta aleatória para fazer a predição de *churn*.

Tabela 1 – Relação entre trabalhos publicados e modelos utilizados

Referências	**Algoritmos**
(CHIANG et al., 2003)	Detecção de Padrões Sequenciais
(JENAMANI et al., 2003)	Semi-Markov
(JONKER et al., 2004)	Algoritmos Genéticos
(LARIVIÈRE; VAN DEN POEL, 2005)	Análise de Sobrevivência
(BUCKINX; VAN DEN POEL, 2005)	Floresta Aleatória
(GOLDSCHMIDT; PASSOS, 2005)	Floresta Aleatória
(LIU; SHIH, 2005)	Filtro por Preferência
(SLOTNICK; SOBEL, 2005)	Semi-Markov
(DE BOCK et al., 2010)	Markov
(BUREZ; VAN DEN POEL, 2007)	Markov e Floresta Aleatória
(KUMAR; RAVI, 2008)	Floresta Aleatória
(COUSSEMENT; VAN DEN POEL, 2008)	Floresta Aleatória
(COUSSEMENT; POEL, 2009)	Regressão Logística, SVM e Floresta Aleatória
(WU, 2009)	Híbrido de Floresta Aleatória e Redes Neurais
(VERBEKE et al., 2012)	Métodos de Conjunto
(ZHANG et al., 2012)	Híbrido envolvendo Aleatória, Regressão logística e Redes Neurais
(COUSSEMENT; DE BOCK, 2013)	Floresta Aleatória

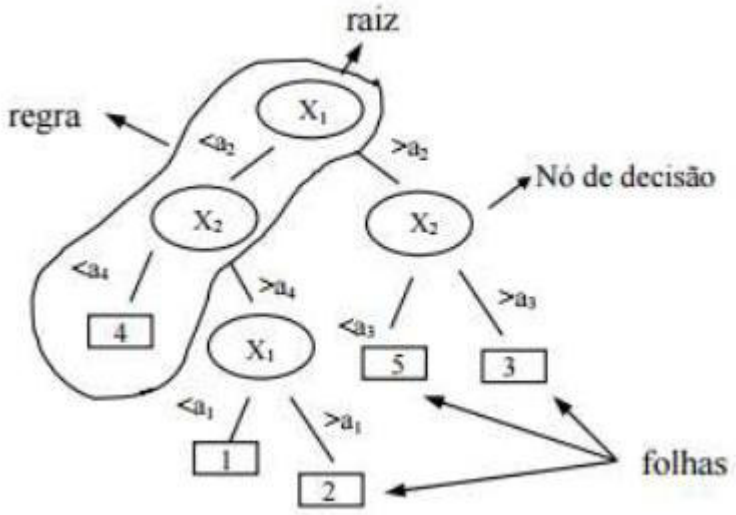
Fonte: PIMENTEL (2019, com adaptações).

3.2.1.1 Floresta Aleatória

Para entender esse modelo, primeiramente é preciso compreender o que são árvores de decisão. Isso porque o *random forest* é a combinação de diversas árvores de decisão.

Árvores de decisão são modelos estatísticos que fazem uso de treinamento supervisionado para prever e classificar dados. Em sua construção é utilizado um conjunto de dados para treinamento formado por entradas e saídas (classes). Esses modelos utilizam a estratégia de dividir para conquistar: um problema complexo é fragmentado em subproblemas de menos complexidade e recursivamente a técnica é aplicada em cada subproblema (GAMA, 2004). A figura 2 representa uma árvore de decisão em que cada nó de decisão contém um teste para algum atributo, o conjunto de ramos são distintos, cada folha está interligada a uma classe e, cada percurso da árvore, da raiz à folha representam uma regra de classificação (GAMA, 2004).

Figura 2 – Representação de uma árvore de decisão



Fonte: GAMA (2004).

O modelo de floresta aleatória é um algoritmo classificador que faz uso do método de árvore de decisão criado por Breiman (2001). Essa técnica efetua a criação de várias árvores de decisão utilizando um subconjunto de atributos selecionados randomicamente a partir do conjunto original, possuindo todos os atributos (NETO, 2014).

Com a fragmentação do conjunto de dados e construção de vários subconjuntos, uma árvore de decisão é construída. A construção das árvores ocorre pela seleção aleatória de atributos a partir dos subconjuntos, em que eles são aplicados nos nós de cada uma das árvores. Após a criação das várias árvores, é possível efetuar a classificação de qual possui melhor lógica e vantagens para a tomada de decisão. Para cada subconjunto é dado um parecer sobre a qual

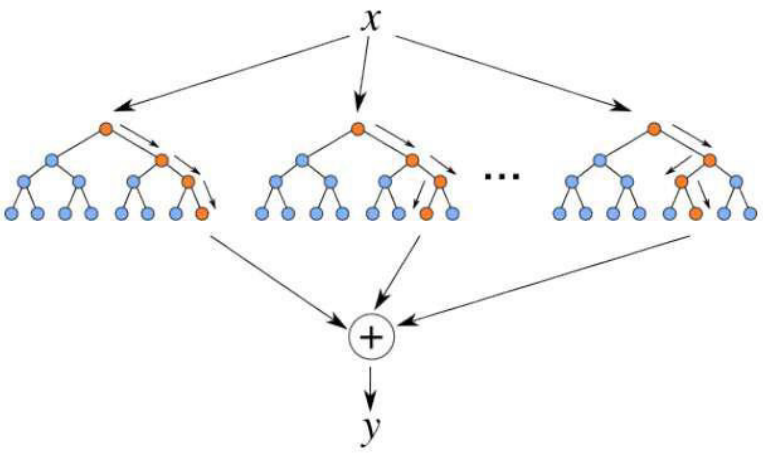
classe o atributo deve pertencer, esse parecer possui um “peso” em que é afetado pela igualdade entre as árvores, ou seja, quanto mais precisa for uma árvore, maior será sua contribuição para a decisão sobre a qual classe o atributo pertence (NETO, 2014). Para exemplificar as árvores de decisão de Breiman, pode-se pensar em uma pessoa que quer jogar tênis. Para isso, são apresentadas algumas características do ambiente, como o aspecto do céu, a temperatura, a humidade e o vento. Cada um desses atributos possui vários valores, a temperatura pode ser amena, fresca ou quente, o aspecto pode ser sol, chuva ou nublado, o vento pode ser fraco ou forte e a humidade pode ser elevado ou normal. Diante das várias combinações, a pessoa irá decidir, baseada nas suas preferências (“pesos”), se irá ou não jogar tênis. Entretanto, outra pessoa irá ter outros tipos de preferências sobre o ambiente para decidir se joga tênis ou não. Sendo assim, o modelo de floresta aleatória utiliza de um conjunto de árvores de decisão com diferentes “pesos” sobre as variáveis (no exemplo proposto seria a opinião das pessoas sobre o ambiente) e classifica o atributo de acordo com a classificação, feita pelas árvores de decisão individuais, que mais se repetiu.

As florestas aleatórias possuem algumas características que as destacam de outras técnicas, sendo algumas delas:

- Algoritmo mais poderoso do que comparado a somente uma árvore de decisão;
- Evita o sobre ajuste (*overfitting)*;
- Possui boa assertividade quando testado e treinando em diferentes conjuntos de dados;
- Menos sensível a ruídos;
- Classificação aleatória das árvores sem intervenção humana.

A figura 3 ilustra o modelo de classificação de florestas aleatórias.

Figura 3 – Ilustração da lógica por trás do algoritmo de floresta aleatória



Fonte: LORENZETT (2016).

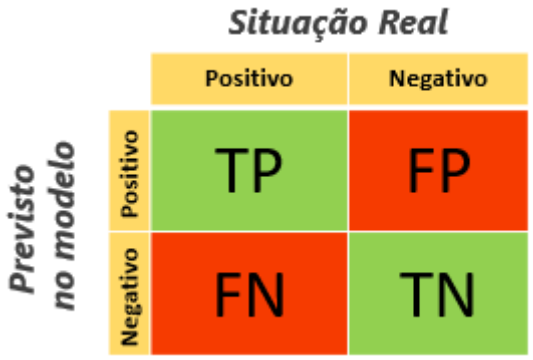
Na imagem, partindo de um elemento *X*, no caso, uma base de dados, foram geradas várias árvores de decisão, em que cada uma produzem várias regras e nelas existe a possibilidade de encontrar novos padrões que poderão ser cruciais na tomada de decisão correta. Após essa etapa, a classificação dada ao atributo será a que mais se repetiu entre o conjunto de árvores geradas.

3.2.2 Métricas de Validação e Performance do Modelo

Conforme Botelho e Tostes (2010), com vistas a minimizar o risco de decisões erradas na classificação dos clientes, diversos métodos estatísticos são utilizados e descrevem a habilidade do modelo de classificação. A avaliação da performance preditiva pode ser feita através da medida da acurácia, precisão, *F1 Score*, dentre outros.

Castro e Braga (2011) ponderam que a métrica mais utilizada é a acurácia, estimada em relação a um dado conjunto de teste. Esse critério é justificável uma vez que se busca a menor probabilidade de erro possível. Porém, ressaltam que uma maneira mais criteriosa de se avaliar um dado classificador é através da distinção dos acertos (ou erros) cometidos para cada classe. Isso é obtido descrevendo o desempenho a partir de uma matriz de confusão, que consiste em uma matriz interpor os dados reais e os valores obtidos no modelo, conforme a figura 4:

Figura 4 – Modelo de matriz de confusão



Fonte: O AUTOR (2021).

Ao longo da diagonal principal (em verde) estão representadas as predições corretas do modelo: verdadeiros positivos (TP) e verdadeiros negativos (TN); os elementos na diagonal secundária representam os erros do modelo: falsos positivos (FP) e falsos negativos (FN).

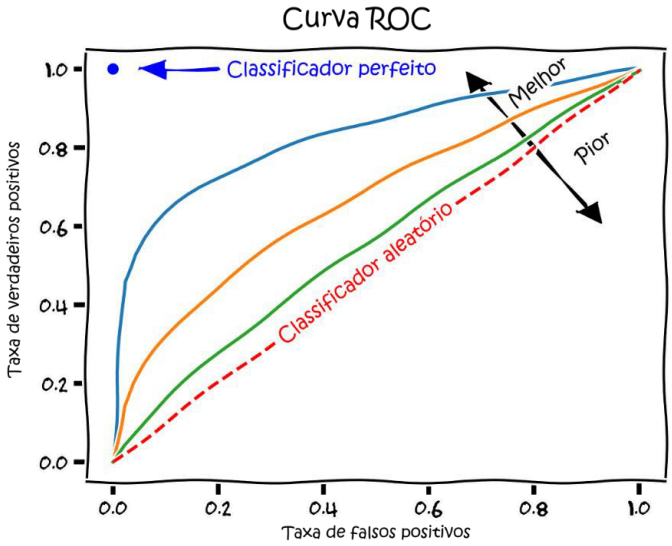
Luque et al. (2019) ressaltam importantes métricas que podem ser obtidas da matriz de confusão. Algumas métricas utilizadas estão descritas a seguir:

- Acurácia: É a medida do total de acertos do modelo pelo total e pode ser obtido pela equação $\frac{TP+TN}{TP+TN+FP+FN}$ (eq.1);
- Precisão: Utilizada para indicar a relação entre as previsões positivas realizadas de maneira assertiva e todas as previsões positivas (incluindo as falsas). É obtido pela equação $\frac{TP}{TP+FP}$ (eq. 2);
- *Recall*: Utilizada para medir a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas. É obtido pela equação $\frac{TP}{TP+FN}$ (eq. 3);
- *F1 Score*: É uma métrica que unifica os resultados da Precisão e do *Recall*. É obtido pela média harmônica entre os dois resultados $\frac{2*\text{Recall}*Precisão}{\text{Recall}+Precisão}$ (eq.4).

Umas das métricas de validação do modelo é o *Hold-out*. Esse método consiste em particionar a base de dados em duas partes, uma para treino e outra para teste. Esse processo é realizado uma única vez e sua vantagem em relação a outros métodos de validação é que o tempo de aprendizagem do modelo é relativamente menor (YADAV et al, 2016).

Outra métrica de validação são as curvas ROC (*Receiver Operating Characteristic*). A curva fornece uma estimativa da capacidade discriminativa do classificador em termos das probabilidades de erros. Assim, controla a fração de exemplos corretamente classificados contra a fração de exemplos incorretamente classificados (CASTRO; BRAGA, 2011). A figura 5 mostra a representação gráfica da curva ROC.

Figura 5 – Modelo de curva ROC



Fonte: Disponível no blog diegomariano.

O classificador (1,0) no plano cartesiano define que o modelo está classificado de forma perfeita, ou seja, ele sempre acertará todas as vezes que precisar definir se é ou não um valor positivo ou negativo.

4 METODOLOGIA E ESTUDO DE CASO

Delimitado o escopo da análise, foi feita a seleção das variáveis que compõem os dados de entrada do modelo preditivo. As variáveis foram escolhidas após um debate entre os analistas dos setores de mercado e sucesso do cliente. Findada a discussão, dez variáveis foram escolhidas para compor o modelo. A tabela 2 discrimina as variáveis selecionadas.

Tabela 2 – Variáveis que compõem o modelo preditivo

Dados do Cliente
Idade do Titular
Valor de Contrato
Forma de Cobrança
Tempo de Base do Cliente
Oferta
Dados de Atendimento
Quantidade de Atendimentos
Dados Financeiros
Número total de faturas
Constância de pagamento de faturas em dias
Média de atraso de pagamento de faturas pagas após vencimento
Representatividade de faturas pagas

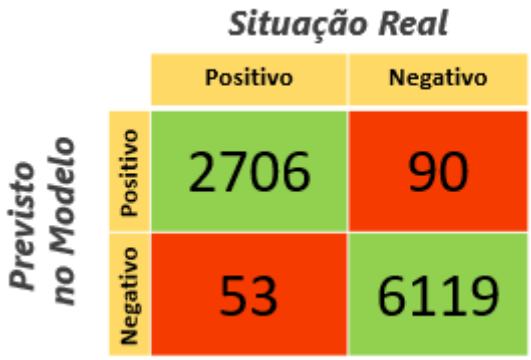
Fonte: O AUTOR (2021).

As etapas de extração, transformação e carga foram feitas todas com o software RStudio®. A amostra de dados gerada possui 29.816 registros, sendo composta pelo *churn* involuntário dos últimos 6 meses e por uma base randômica de clientes ativos. Os dados foram divididos, mutuamente exclusivos, em 70% para aprendizado e 30% para teste. Tal disposição dos dados foi feita para validação do modelo através do método *Hold-out*. O modelo de floresta aleatória utilizou 500 árvores de decisão para avaliar cada cliente na base de aprendizado.

5 RESULTADOS

A matriz de confusão obtida através da aplicação do modelo construído na base de treino ter sido aplicado na base de teste está ilustrada na figura 6.

Figura 6 – Matriz de confusão do modelo aplicado



Fonte: O AUTOR (2021).

O modelo classificou de forma errônea 143 dados, sendo 90 do tipo falso negativo e 53 do tipo falso positivo. A tabela 3 resume as estatísticas de performance do modelo obtidas através da matriz de confusão.

Tabela 3 – Resultado das estatísticas de performance

Métrica	Fórmula	Resultado
Acurácia	TP+TN/TP+FP+TN+FN	0,9840
Precisão	TP/TP+FP	0,9678
Recall	TP/TP+FN	0,9807
F1 – Score	Média Harmônica (R,P)	0,9742

Fonte: O AUTOR (2021).

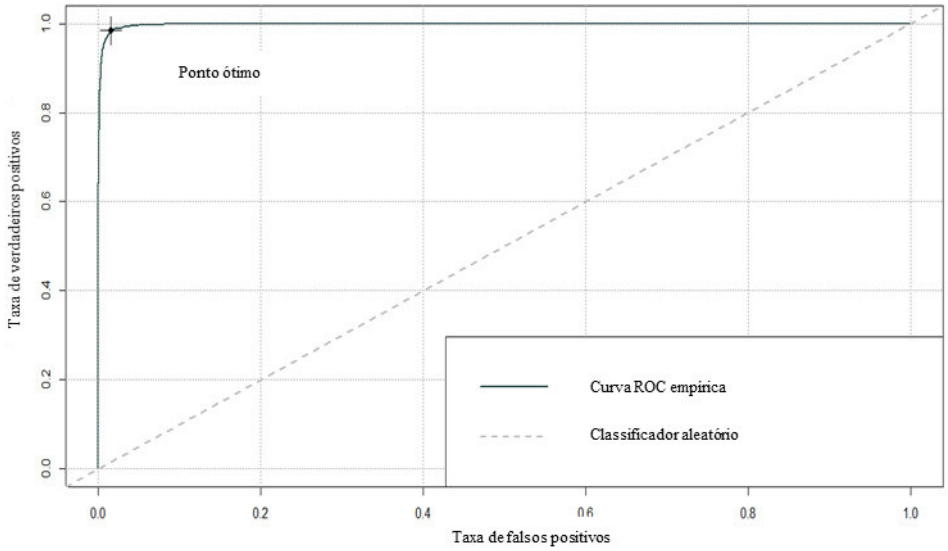
Pode-se fazer as seguintes inferências baseado nos resultados das estáticas de performance:

- A acurácia em 98% permite afirmar que para 100 classificações que o modelo fez, em 98 vezes o fez de forma correta;

- A precisão em 97% garante que de 100 resultados que o modelo classificou como verdadeiro, 97 das classificações realmente eram verdadeiras;
- O *recall* em 98% garante que para 100 resultados positivos, o modelo classificará de forma correta 98 deles;
- O *F1 Score* unifica os resultados da precisão e do *recall*, a métrica ficou muito próximo de 100% o que reforça a qualidade do modelo utilizado.

A figura 7 apresenta a curva ROC do modelo construído.

Figura 7 – Curva ROC do modelo construído



Fonte: O AUTOR (2021).

O modelo performa muito próxima ao classificador perfeito, o que garante que as variáveis selecionadas para alimentar o modelo possuem correlação e causalidade com o *churn*. Tanto a curva ROC quanto as métricas de performance corroboram para afirmar que o modelo possui uma grande força preditiva.

6 CONCLUSÃO

A escolha do modelo de floresta aleatória para a classificação dos clientes em ser ou não *churn* mostrou-se bastante assertiva, uma vez que as métricas de validação e de performance apresentaram resultados satisfatórios. Dessa forma, o objetivo do presente trabalho foi atingido, pois a classificação dos clientes foi feita de forma bastante precisa. O modelo foi implementando como ferramenta na empresa e medidas de retenção de clientes estão sendo aplicadas nos clientes em que o modelo classifica como *churn*.

7 TRABALHOS FUTUROS

- Comparativo do modelo de floresta aleatória com outros modelos preditivos, como o SVM – *Support Vector Machine* e a regressão logística, verificando a performance dos modelos;
- Utilização do modelo de floresta aleatória em uma aplicação na metalurgia, como classificação de defeitos em fundidos.

REFERÊNCIAS

A. PAYNE, P. F. A strategic framework for customer relationship management. Journal of Marketing Research, v. 69, p. 167—176, 2005.

BALLE, B.; CASAS, B.; CATARINEU, A. The Architecture of a Churn Prediction System Based on Stream Mining, Lsi.upc.edu, 2011.

BOTELHO, Delane; TOSTES, Frederico Damian. Modelagem de probabilidade de churn. Revista de Administração de Empresas, São Paulo, v. 4, n. 396, 2010.

BREIMAN, L. Random Forests, Machine Learning, Vol. 45, pp. 5 – 32, 2001.

BUREZ, J.; VAN DEN POEL, D. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications, v. 32, p. 277–288, 2007.

CASTRO, C.L. de; BRAGA, A. P; Supervised learning with imbalanced data sets: an overview. Sba: Controle & Automação Sociedade Brasileira de Automatica, Campinas, SP, n. 5, p. 441, 2011.

CHEESEMAN, P.; J. STUTZ. Bayesian classification (Autoclass): Theory and results advances in knowledge discovery and data mining. NASA. Disponível em: <https://www.researchgate.net/publication/2659256_Bayesian_ClassificationAutoClassTheory_and_Results> Acesso em: 11 maio de 2021.

COUSSEMENT, K.; POEL, D. V. D. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, v. 36, p. 6127–6134, 2009.

DANTAS, Guilherme Vieira. Utilização de classificador random forest na detecção e previsão de falhas em máquinas rotativas. Monografia (Graduação em Engenharia Eletrônica e de Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

Diegomariano, Métricas de avaliação em machine learning. Disponível em: < https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/>. Acesso em: 15 de agosto de 2021.

GAMA, J. Notas de aula. USP, 2004.

GARCÍA, D. L.; NEBOT; VELLIDO, A. Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems, v. 51, p. 719–774, 2017.

GOLDSCHMIDT, E. B.; PASSOS, E. Data Mining: Um Guia Prático. 2nd. ed. [S.l.: s.n.], 2015. 296 p.

GUEDES, Érick Victor De Oliveira. Aplicação de Soft Sensor Baseado em Redes Neurais Artificiais e Random Forest para Predição em Tempo Real do Teor de Ferro no Concentrado da Flotação de Minério de Ferro. Dissertação (Mestrado em Engenharia de Controle e Automação) – Universidade Federal de Ouro Preto, Ouro Preto, 2020.

HADDEN, J.; TIWARI, A.; ROY, R.; RUTA, D. Computer assisted customer churn management: State-of-the-art and future trends. Computers and Operations Research, v. 34, p. 2902–2917, 2007.

KOTSIANTIS, S. B. "Supervised machine learning: a review of classification techniques, Informatica 31, 249–268, 2007. Disponível em: <https://datajobs.com/data-science repo/Supervised-Learning- [SB-Kotsiantis].pdf> Acesso em: 10 de maio de 2021.

LORENZETT, C. D. C; TELÖCKEN, A. V. Estudo comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de decisão. Curso de Ciência da Computação. Universidade de Cruz Alta (UNICRUZ). Rio Grande do Sul, 2016.

LUQUE, A. et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, [s. l.], v. 91, p. 216–231, 2019.

MARTÍNEZ, Estela. Machine learning algorithms for the prediction of non-metallic inclusions in steel wires for tire reinforcement. 2019.

MONARD, M.C.; BARANAUSKAS, J.A. Conceitos sobre aprendizado de máquina. In: Rezende, S.O. (Ed.). Sistemas Inteligentes: Fundamentos e aplicações. São Carlos, 2003. p.89-114. cap. 4.

NETO, C. Di G. Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras. INPE, São José dos Campos, 2014.

PEREIRA, Kaléo Dias. Aprendizagem de máquina e técnicas multivariadas no estudo da qualidade do carvão vegetal. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, 2019.

PIMENTEL, Thiago Paiva. Predição de churn baseada em detecção de padrões sequenciais e análise de sentimentos sobre as interações de clientes no CRM. Dissertação (Mestrado em Ciências em Sistemas e Computação) - Instituto Militar de Engenharia, Rio de Janeiro, 2019.

RAUTIO, A. Churn prediction in saas using machine learning, 2019.

ROJAS, R. Neural Nerworks: A systematic Introduction. Springer-Verlag, Berlin, 1996. Disponível em: <https://page.mi.fu-berlin.de/rojas/neural/neuron.pdf> Acesso em: 11 de maio de 2021.

YADAV, Sanjay; SHUKLA, Sanyam. Analysis of k-Fold Cross-Validation over Hold Out Validation on Colossal Datasets for Quality Classification. In: Advanced Computing (IACC), 2016 IEEE 6th International Conference on. IEEE, 2016. p. 78-83.