UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO TECNOLÓGICO

COLEGIADO DO CURSO DE ENGENHARIA DE COMPUTAÇÃO

Luiza Batista Laquini

Caracterização e predição da (in)satisfação de clientes de uma operadora de telefonia móvel

Vitória, ES

2023

Luiza Batista Laquini

Caracterização e predição da (in)satisfação de clientes de uma operadora de telefonia móvel

Monografia apresentada ao Colegiado do Curso de Engenharia de Computação do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito para obtenção do grau de Bacharel em Engenharia de Computação.

Universidade Federal do Espírito Santo (UFES)

Centro Tecnológico

Colegiado do Curso de Engenharia de Computação

Orientador: Prof. Vinícius Fernandes Soares Mota

Vitória, ES

2023

Luiza Batista Laquini
Caracterização e predição da (in)satisfação de clientes de uma operadora de telefonia móvel/ Luiza Batista Laquini. – Vitória, ES, 2023-
65 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Vinícius Fernandes Soares Mota

Monografia (PG) – Universidade Federal do Espírito Santo (UFES)
Centro Tecnológico
Colegiado do Curso de Engenharia de Computação, 2023.

1. NPS. 2. Telefonia. 3. Caracterização. 4. Predição. I. Luiza Batista Laquini. II. Universidade Federal do Espírito Santo. IV. Caracterização e predição da (in)satisfação de clientes de uma operadora de telefonia móvel

CDU 02:141:005.7

Luiza Batista Laquini

Caracterização e predição da (in)satisfação de clientes de uma operadora de telefonia móvel

Monografia apresentada ao Colegiado do Curso de Engenharia de Computação do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito para obtenção do grau de Bacharel em Engenharia de Computação.

Trabalho aprovado. Vitória, ES, 15 de dezembro de 2023:

Prof. Vinícius Fernandes Soares Mota
Orientador

Prof. Giovanni Ventorim Comarela
Convidado 1

Vitor Fontana Zanotelli
Convidado 2

Vitória, ES
2023

Dedico este trabalho a todos os que me ajudaram ao longo desta caminhada.

Agradecimentos

Em primeiro lugar, a Deus, que me deu forças nos momentos de crise e iluminou o meu caminho durante todos esses anos de estudo.

Ao meu noivo, André, que foi quem esteve mais próximo de mim durante todos esses anos, comemorando os momentos de vitória e me incentivando a me reerguer quando algo não saía conforme o planejado.

Aos meus pais, Sérgio e Eli, aos meus irmãos, Sérgio Júnior e Raphael e aos demais membros da família que sempre me apoiaram na escolha do curso e compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho e de todas as provas e projetos da Universidade.

Ao meu professor e orientador, Vinícius Mota, que me propôs esse projeto de graduação, o qual me identifiquei tanto e cujo me permitiu crescer e abrir portas na carreira profissional.

Ao professor Giovanni Comarela, que também acompanhou diretamente esse projeto e lecionou disciplinas essenciais para o desenvolvimento do mesmo.

A equipe TIM, que trabalhou junto comigo dando direcionamentos e interpretando resultados. Em especial, aos colegas Vitor Zanotelli, Wadham Entringer e Iran Ribeiro, que se reuniram diversas vezes comigo, tendo contribuído muito para o meu avanço e conhecimento.

A todo o corpo docente do departamento de informática da UFES que contribuiu para a minha formação como Engenheira da Computação. Em especial, às mulheres, que sempre foram de grande inspiração em meio a uma área de conhecimento formada majoritariamente por homens.

Aos meus colegas de curso, com quem convivi intensamente durante os últimos anos, pelo companheirismo e pela troca de experiências que me permitiram crescer não só como profissional, mas também como pessoa.

Aos meus gatinhos, Simba e Mia, que aqueceram o meu colo nas madrugadas de estudo, sem sair de perto até que eu houvesse finalizado todas as minhas tarefas. Seus olhares me apoiavam e eu nunca me senti sozinha.

Resumo

O projeto em questão aborda a caracterização e predição da satisfação dos clientes de uma operadora de telefonia móvel. A caracterização será realizada por meio da análise do Net Promoter Score (NPS) aplicado a diferentes características, proporcionando uma compreensão abrangente dos fatores que influenciam a satisfação do cliente.

Além disso, a predição da satisfação dos clientes será abordada por meio do treinamento de algoritmos de aprendizado de máquina. Essa abordagem permitirá antecipar a satisfação ou insatisfação futura com base em padrões históricos identificados nas variáveis analisadas.

A satisfação do cliente é essencial para o sucesso das operadoras de telefonia móvel, e a capacidade de antecipar e compreender as razões por trás da insatisfação permite que as empresas adotem estratégias eficazes para retenção de clientes. Utilizando conjuntos de dados fornecidos pela operadora TIM, o estudo emprega técnicas de análise exploratória e aprendizado de máquina para identificar padrões e correlações nas características de uso dos usuários.

As etapas envolvidas são: pré-processamento de dados, análise exploratória, treinamento de algoritmos, avaliação de desempenho e, por fim, a apresentação de insights e recomendações para a operadora de telefonia móvel. A integração do NPS com métodos avançados de predição visa fornecer estudos que se complementam para entender, antecipar e melhorar a satisfação do cliente, contribuindo assim para o sucesso e crescimento sustentável do negócio.

Palavras-chave: NPS; Telefonia; Caracterização; Predição.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC *Area Under the Curve*

CART *Classification and Regression Trees*

CSAT *Customer Satisfaction Score*

CSV *Comma Separated Values*

CX *Customer Experience*

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*

GBM *Gradient Boosted Machine*

HDFS *Hadoop Distributed File System*

IDH Índice de Desenvolvimento Humano

IQR *Interquartile Range*

KNN *K-Nearest Neighbors*

ML *Machine Learning*

NPS *Net Promoter Score*

ROC *Receiver Operating Characteristic*

SHAP *SHapley Additive exPlanations*

SVM *Support Vector Machine*

XGBOOST *Extreme Gradient Boosting*

Sumário

1 Apresentação

O primeiro capítulo desta monografia desempenha um papel fundamental ao estabelecer as bases essenciais para a compreensão do escopo, propósito, motivação e abordagem metodológica do trabalho.

1.1 Introdução

Nos últimos anos, os serviços de telefonia móvel desempenharam um papel vital na comunicação e interação social. Com os avanços tecnológicos e a crescente demanda por serviços móveis, as operadoras de telefonia são desafiadas a garantir a satisfação do cliente. A satisfação do cliente é um indicador importante para avaliar a qualidade do serviço prestado e a fidelização do cliente a um determinado serviço.

Um importante prognóstico relacionado à satisfação do cliente é a identificação de detratores, ou seja, usuários insatisfeitos. Esses têm maior probabilidade de abandonar o serviço da operadora e prejudicar a imagem da empresa. A capacidade de identificar e entender os detratores permite que as operadoras adotem estratégias eficazes de retenção de clientes, forneçam soluções personalizadas para resolver seus problemas e melhorem a qualidade do serviço. Da mesma forma, também é interessante entender os clientes promotores (satisfeitos), pois isso permite que a empresa adote uma abordagem proativa para melhorar áreas específicas que são consistentemente apreciadas pelos usuários. Isso não apenas aumenta a satisfação dos clientes existentes, mas também contribui para a aquisição de novos. A imagem positiva criada pelos promotores é uma ferramenta poderosa de marketing, atraindo novos usuários que procuram uma experiência positiva com os serviços de telefonia móvel.

O presente projeto tem como objetivo entender os perfis dos clientes, identificando padrões e correlações relevantes utilizando-se de técnicas de análise exploratória e cálculos da métrica *Net Promoter Score* para avaliação da satisfação dos mesmos. Essa métrica é amplamente discutida na Seção 2.1, mas basicamente se trata do percentual de detratores subtraído do percentual de promotores. Além disso, serão desenvolvidos modelos de *machine learning* capazes de predizer detratores com base nas características de uso dos usuários de serviços de telefonia móvel.

Esse estudo será baseado em conjuntos de dados fornecidos pela empresa de telefonia TIM. Esses dados contêm respostas à pesquisa de satisfação feita pela empresa, propriedades como DDD do cliente e plano contratado, estatísticas de consumo como total de ligações e volumes de dados, entre outras informações que serão discutidas no

decorrer desta monografia. Vale mencionar que a pesquisa de satisfação foi realizada em duas etapas (perguntas). A primeira delas pediu para o cliente atribuir uma nota de zero a dez para o serviço da TIM e, a segunda, pediu a justificativa para a nota atribuída, por meio de opções prontas.

O escopo do projeto envolve etapas de pré-processamento dos dados fornecidos, análise exploratória, modelagem e predição, avaliação e validação e, por fim, geração de insights e recomendações à operadora de telefonia móvel. A aplicação dessas técnicas permitirá à operadora tomar medidas proativas para melhorar a satisfação do cliente e a retenção, contribuindo para o sucesso e o crescimento do negócio.

1.2 Motivação e Justificativa

A escolha do tema decorre do estado atual do mercado de telecomunicações. À medida que a concorrência se intensifica e a tecnologia continua a evoluir, as operadoras móveis enfrentam desafios significativos para manter a satisfação e a fidelidade do cliente. A identificação precoce de usuários insatisfeitos e propensos a abandonar o serviço tornou-se fundamental para aprimorar as estratégias de retenção de clientes e garantir a sustentabilidade do negócio.

Nesse sentido, acredita-se que os resultados obtidos contribuirão para o aperfeiçoamento das estratégias de retenção de clientes e para a melhoria dos serviços prestados pelas operadoras, beneficiando, assim, tanto a empresa quanto os usuários finais.

1.3 Objetivos

1.3.1 Objetivo Geral

Identificar os perfis de satisfação dos clientes e predizer um cliente que possui maior probabilidade de ser um detrator.

1.3.2 Objetivos Específicos

- Analisar os dados disponibilizados de forma detalhada visando identificar padrões e tendências.
- Investigar os indicadores que estão associados à insatisfação do cliente e ao potencial de se tornarem detratores, com um possível abandono dos serviços de telefonia móvel.
- Desenvolver modelos preditivos capazes de apontar os clientes com maior probabilidade de serem detratores.

1.4 Método de Desenvolvimento do Trabalho

A primeira etapa é a de pré-processamento. Nessa fase, foram aplicadas técnicas de limpeza, integração, transformação e redução de dimensionalidade para preparar o conjunto de dados. Foram removidos dados duplicados, tratados valores faltantes ou discrepantes e ajustados possíveis erros de formatação.

Em seguida, foram utilizadas técnicas de visualização de dados. Essa análise inicial forneceu os primeiros insights sobre os dados e orientou as etapas subsequentes da análise.

Em terceiro lugar, foram aplicadas técnicas de análise exploratória para identificar e levantar hipóteses sobre o perfil dos usuários detratores e dos promotores com base no cálculo do *Net Promoter Score* (NPS).

Não obstante, foram desenvolvidos modelos de classificação utilizando técnicas de aprendizado de máquina. Os modelos foram treinados utilizando diferentes separações dos dados que estão bem detalhadas na Seção 4.5. Foram explorados diferentes algoritmos, como KNN, Regressão Logística, Árvore de Decisão, Random Forest e *Extreme Gradient Boosting*. Os que apresentaram melhor desempenho inicial passaram por uma etapa a mais que é a otimização de hiperparâmetros.

Por fim, os modelos desenvolvidos foram avaliados e validados com base nas métricas de desempenho: acurácia, precisão, *recall* e *f1-score*. Com base nos resultados obtidos, foram discutidos *insights* relevantes para a operadora de telefonia móvel e foram deixadas sugestões para trabalhos futuros.

2 Fundamentação Teórica e Tecnologias Utilizadas

A seguir, apresentamos os aspectos relativos ao conteúdo teórico necessário para compreensão do projeto, ao conteúdo do material bibliográfico consultado e às tecnologias que irão auxiliar no desenvolvimento do trabalho em busca de uma solução para o problema abordado.

2.1 *Net Promoter Score*

Uma métrica importante de satisfação do cliente que se tornou muito popular nos últimos anos e que é uma das bases da análise deste trabalho é o *Net Promoter Score* (NPS). Ela foi desenvolvida por Fred Reichheld, consultor de gestão e autor do livro *"The Ultimate Question"*. Ele introduziu o conceito em um artigo publicado na *Harvard Business Review* em 2003, intitulado *"The One Number You Need to Grow"*.

O cálculo do NPS foi proposto por (REICHHELD, 2003) e é baseado em uma pergunta simples que é feita aos clientes:

"Em uma escala de 0 a 10, o quanto você recomendaria nossa empresa/-produto/serviço para um amigo ou colega?".

Dessa maneira, os clientes têm a oportunidade de avaliar o negócio e atribuir notas em escala. A partir das respostas, é possível classificá-los nas seguintes categorias:

(1) Promotores: aqueles que estão completamente satisfeitos com o serviço oferecido, estes responderam à pesquisa com as notas 9 e 10; (2) Neutros: clientes que não estão satisfeitos nem insatisfeitos, ou seja, indiferentes. Eles atribuíram as notas 7 e 8; (3) Detratores: consumidores que não estão nada satisfeitos e tiveram uma experiência muito negativa com o negócio. Eles responderam à pesquisa com notas de 0 a 6.

O resultado final é feito subtraindo a porcentagem de clientes detratores da porcentagem de clientes promotores. O resultado pode variar de -100 a 100. Um NPS positivo indica que a maioria dos clientes é promotor, enquanto um NPS negativo indica que a maioria dos clientes é detrator. Essa análise pode levantar hipóteses relevantes para o encaminhamento do projeto.

Vale ressaltar que cada empresa tem alguma flexibilidade para definir suas próprias faixas de notas e critérios de classificação de detratores, neutros e promotores. No entanto, a escala de 0 a 10 e a divisão básica entre as categorias que foram estabelecidas por Reichheld são amplamente seguidas na aplicação do NPS.

2.2 Pré-Processamento dos Dados

Essa etapa é dependente do tipo dos dados, do problema em questão e dos requisitos específicos da análise. Ela visa preparar os dados brutos coletados, eliminando inconsistências, reduzindo o ruído e deixando os dados no formato adequado para aplicação de modelos preditivos. Isso porque uma tabela de dados real, na grande maioria das vezes, possui dados ausentes ou infinitos, erros de formatação, dados irrelevantes ou duplicados, entre outros casos. A escolha da técnica apropriada para realizar o pré-processamento de dados pode influenciar diretamente no desempenho dos modelos e na interpretabilidade dos dados.

2.2.1 Limpeza de Dados

A etapa de limpeza de dados tem como objetivo identificar e lidar com dados inconsistentes, incorretos ou irrelevantes. Durante essa fase, são realizadas atividades como remoção de duplicatas, padronização de formatos de dados e tratamento de dados ausentes ou infinitos. Além disso, é importante verificar a integridade dos dados, identificando e tratando *outliers* (valores discrepantes).

Remoção de Duplicatas

A remoção de duplicatas é uma tarefa importante para garantir a qualidade dos dados. Ela pode ser realizada por meio da comparação dos registros com base em um ou mais atributos-chave, como número de telefone ou identificador de cliente. Os registros duplicados podem ser eliminados ou fundidos, dependendo da necessidade do estudo.

Padronização dos Formatos

A padronização dos formatos de dados é necessária para garantir que os dados estejam consistentes e possam ser corretamente interpretados. Isso pode envolver a conversão de dados para um formato específico, como transformar números de telefone em um formato padronizado ou converter datas para um formato comum.

Tratamento dos Dados Ausentes ou Infinitos

Dados ausentes ou infinitos são uma ocorrência comum em conjuntos de dados reais. O tratamento adequado desses valores é essencial para evitar viés e distorções nos resultados. Existem várias abordagens para lidar com valores ausentes ou infinitos, incluindo exclusão de registros, substituição por valores médios ou estimativas baseadas em outros atributos.

A exclusão de registros com valores ausentes pode ser uma opção viável se a quantidade de registros faltantes for pequena em relação ao tamanho do conjunto de dados.

No entanto, essa abordagem pode levar à perda de informações importantes se os registros excluídos contiverem informações valiosas.

A substituição por valores médios é uma técnica comumente usada, onde os valores ausentes são substituídos pela média dos valores existentes para o mesmo atributo. Essa abordagem assume que os valores ausentes são semelhantes aos valores observados.

Outras técnicas mais avançadas podem ser utilizadas, como a imputação de dados por meio de regressão, onde um modelo é construído para estimar os valores ausentes com base em outros atributos relevantes.

2.2.2 *Encoding*

Encoding é o processo de converter dados de uma forma para outra. Essa etapa é necessária quando há colunas de tipos que não são interpretáveis pelos modelos na base de dados. Muitos algoritmos de *machine learning* suportam apenas dados numéricos como entrada. Portanto, variáveis categóricas frequentemente precisam ser convertidas para valores numéricos antes de serem fornecidas a esses algoritmos.

A escolha entre as diferentes técnicas depende do tipo de dado, das características específicas do conjunto de dados e dos requisitos do modelo. Cada abordagem tem vantagens e limitações, e a escolha certa depende do contexto da análise.

O algoritmo *One Hot Encoder*

O algoritmo mais amplamente utilizado para a tranformação de dados categóricos em binários, especialmente se você tiver muitas categorias é o *One Hot Encoder*. Ele é eficaz para representar categorias sem atribuir significado ordinal às mesmas. Se a ordem entre as categorias é relevante, é importante considerar outras técnicas de codificação que preservem essa informação.

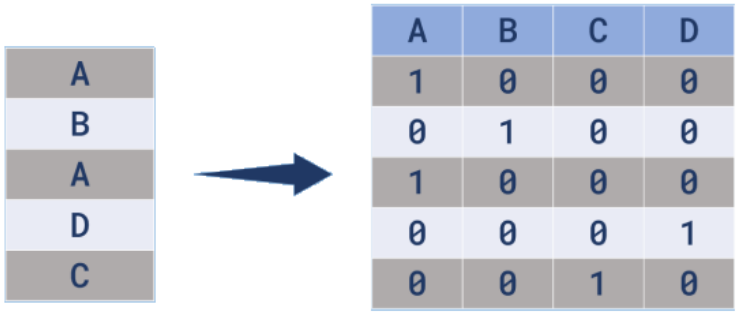

Figura 1 – Tranformação de colunas com o algoritmo *One Hot Encoder*. Imagem disponível em: <https://www.alura.com.br/artigos/get-dummies-vs-onehotencoder-qual-metodo-escolher>

Ele funciona criando uma nova coluna para cada categoria existente na base de dados, onde será atribuido o valor 1 indicando a ocorrência daquela categoria. Essa transformação é ilustrada na Figura 1.

2.2.3 Normalização dos Dados

A normalização dos dados é uma etapa importante para garantir que diferentes variáveis estejam na mesma escala e sejam considerados de maneira equilibrada - antes de serem submetidos à algoritmos de aprendizado de máquina - evitando assim que algumas variáveis dominem a análise devido a seus valores numéricos. É importante observar que a normalização dos dados é realizada apenas nos atributos de entrada, não afetando a variável alvo ou o resultado da análise em si. Existem várias técnicas de normalização para diferentes padrões de dados, mas para esse projeto vamos fazer uma breve introdução de três algoritmos: o *Standard Scaler*, o *MinMaxScaler* e o *MaxAbsScaler*.

Um estudo de (RAJU et al., 2020), descreve a melhoria nas precisões preditivas com a ajuda de técnicas de normalização. Vários critérios necessários para atingir essa normalização de dados também são descritos. Para isso, os autores utilizam sete técnicas de padronização diferentes (incluindo as três que serão abordadas à seguir) em cima de alguns algorimos de aprendizado de máquina e comparam os valores de acurácia obtidos antes e depois da aplicação dos mesmos.

O algoritmo *Standard Scaler*

É recomendado quando os dados têm distribuição normal ou aproximadamente normal. Ele transforma os dados para que tenham média zero e desvio padrão igual a um. Isso significa que, após a aplicação, os dados terão uma distribuição aproximadamente gaussiana, o que é útil para algoritmos sensíveis à escala dos dados, como métodos baseados em distância e regressão linear. Todos esses detalhes foram bem descritos por (PEDREGOSA et al., 2011). Uma síntese do funcionamento é abordada a seguir.

A fórmula matemática para a transformação dos dados, que opera em cima de cada recuso, utilizando o *Standard Scaler* é dada por:

$$x_{standardized} = \frac{x - \mu}{\sigma} \tag{2.1}$$

Onde:

- $x_{stardardized}$ é o valor transformado do atributo;
- x é o valor original do atributo;
- μ é a média dos valores do atributo;

- σ é o desvio padrão dos valores do atributo.

Dessa forma, o algoritmo opera calculando a média (*mean*) e o desvio padrão (*standard deviation*) dos valores em cada atributo. Em seguida, subtrai-se a média de cada valor e divide-se pelo desvio padrão. Essa transformação garante que os dados fiquem centrados em torno de zero e tenham uma escala comparável.

O algoritmo *MinMaxScaler*

É adequado quando os dados não seguem uma distribuição normal e têm valores limitados em um intervalo específico. Ele mantém a forma geral da distribuição dos dados, escalando-os para um intervalo específico, geralmente entre 0 e 1. Isso é útil para algoritmos sensíveis à escala e que requerem que os dados estejam em intervalos, como algoritmos de otimização baseados em gradiente.

A fórmula matemática para a transformação dos dados, que opera em cima de cada recuso, utilizando o *MinMaxScaler* é dada por:

$$x_{scaled} = \frac{x - min}{max - min} \tag{2.2}$$

Onde:

- x_{scaled} é o valor escalado resultante após a aplicação da normalização.
- x é o valor original do atributo;
- min é o valor mínimo desse atributo no conjunto de dados;
- max é o valor máximo desse atributo no conjunto de dados;

Assim, o algoritmo garante que todos os valores sejam ajustados para um novo intervalo, preservando a relação de ordem e a distribuição relativa dos dados.

No entanto, é importante notar que o *MinMaxScaler* é sensível a *outliers* (valores discrepantes), pois os valores mínimos e máximos são determinados pelo conjunto de dados. Portanto, se houver *outliers* significativos, eles podem distorcer o escalonamento dos dados. Nesse caso, é extremamente importante considerar técnicas de detecção e tratamento de *outliers* antes de aplicá-lo.

O algoritmo *MaxAbsScaler*

É útil quando os dados possuem valores dispersos, como em dados esparsos ou em dados binários. O algoritmo escala os dados de forma que o valor absoluto máximo de cada atributo seja 1, preservando a esparsidade dos dados. Isso é adequado para algoritmos que

não assumem nenhuma distribuição específica nos dados, como algoritmos de aprendizado baseados em árvores de decisão ou métodos de redução de dimensionalidade.

A fórmula matemática para a transformação dos dados, que opera em cima de cada variável, utilizando o *MaxAbsScaler* é dada por:

$$x\_scaled = \frac{x}{max\_abs} \tag{2.3}$$

Onde:

- x_scaled é o valor escalado resultante após a aplicação da normalização.
- x é o valor original do atributo.
- max_abs é o valor absoluto máximo desse atributo no conjunto de dados.

De tal maneira, o algoritmo divide cada valor pelo valor absoluto máximo desse atributo no conjunto de dados, garantindo que todos os valores escalados estejam dentro do intervalo [-1, 1]. É importante notar que isso pode não ser apropriado ao lidar com dados que tenham valores extremamente grandes, pois a escala final será muito pequena. No entanto, essa abordagem preserva a relação de ordem e a esparsidade dos dados.

Ao contrário do *MinMaxScaler*, uma vantagem do *MaxAbsScaler* é que ele é menos sensível a *outliers*, pois apenas o valor absoluto máximo do atributo é considerado. Todavia, é importante ressaltar que ele não preserva a forma geral da distribuição dos dados, apenas garante que todos os valores escalados estejam dentro do intervalo [-1, 1].

2.3 Correlação Linear

Quando falamos de análise de dados, muitas vezes precisamos entender a associação entre duas ou mais variáveis. A análise de correlação nesse contexto é uma forma descritiva que mede se - e quanto - existe dependência linear entre variáveis, ou seja, quanto uma variável interfere na outra linearmente. O grau dessa relação é medido através de coeficientes. Para este projeto, por se tratar do padrão amplamente utilizado em conjuntos de dados como os de telefonia, vamos nos concentrar no coeficiente de Pearson, que possui um guia publicado por (DOE; SMITH, 2022). Segundo os autores, o coeficiente de Pearson mede o grau de correlação através do cálculo de direção positiva ou negativa e assume apenas valores entre -1 e 1. Para duas variáveis, x e y, o cálculo é feito da seguinte maneira:

$$\rho_{XY} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{(n-1)\sigma_X \sigma_Y} \tag{2.4}$$

Onde:

- ρ_{XY} é o coeficiente de Pearson
- $\bar{x}$ é o valor médio da amostra para a variável x
- $\bar{y}$ é o valor médio da amostra para a variável y
- n é o número de observações
- σ_X é o desvio padrão da amostra para a variável x
- σ_Y é o desvio padrão da amostra para a variável y

A análise de correlação retornará três casos possíveis: (1) correlação positiva, quando o resultado se aproxima de 1; (2) correlação negativa, quando o resultado se aproxima de -1; (3) nenhuma correlação, quando o resultado é proximo de 0 (zero). A presença de correlações lineares entre as *features* de um projeto deve ser logo verificada, pois facilita a compreensão e simplifica bastante a análise e a predição uma vez que podemos entender melhor os dados como um todo.

Apesar disso, como o próprio nome sugere, essas correlações são lineares. Variáveis podem ter relações não lineares que esse cálculo não irá identificar. Dessa forma, tornam-se necessários os algoritmos de aprendizado de máquina para identificar padrões não lineares nos dados.

2.4 Aprendizado de Máquina

O Aprendizado de Máquina, do inglês *Machine Learning* (ML), envolve o desenvolvimento de algoritmos e técnicas que permitem aos computadores aprender padrões de uma base de dados e tomar decisões para novas entradas.

Nesta seção, apresentaremos uma visão geral das diferentes categorias de problemas que existem, incluindo algoritmos que podem vir a serem utilizados. Vale ressaltar que os conhecimentos aqui apresentados foram retirados do livro (BISHOP, 2006), uma obra amplamente reconhecida no campo, que fornece uma base sólida para compreensão dos conceitos e algoritmos discutidos.

2.4.1 Categorias

O aprendizado de máquina pode ser classificado em algumas categorias. Cada categoria possui diferentes abordagens e técnicas que podem ser aplicadas em uma ampla gama de problemas e domínios. As duas principais categorias de aprendizado relevantes para este trabalho são:

- Aprendizado Supervisionado: Nesse tipo de aprendizado, o modelo é treinado usando um conjunto de dados rotulados, em que cada exemplo possui uma entrada e a saída correspondente. O objetivo é fazer o modelo aprender a mapear as entradas para as saídas corretas.

- Aprendizado Não Supervisionado: No aprendizado não supervisionado, o modelo é treinado em um conjunto de dados não rotulados, onde não há informações sobre as saídas desejadas. O objetivo é encontrar padrões, estrutura e informações úteis nos dados.

Tradicionalmente, no aprendizado supervisionado temos problemas de classificação ou de regressão. Já no aprendizado não supervisionado temos problemas de agrupamento/-clusterização. A seguir, para cada tipo de problema serão abordados os algoritmos mais utilizados.

2.4.2 Algoritmos de Classificação

Um problema de classificação em Aprendizado de Máquina refere-se a um tipo de tarefa em que o objetivo é atribuir rótulos ou categorias a instâncias de dados com base em suas características. O objetivo é encontrar um modelo ou algoritmo que seja capaz de aprender a relação entre os atributos das instâncias e suas respectivas classes, a fim de realizar previsões ou classificações corretas para novas instâncias não vistas anteriormente.

Alguns dos algoritmos de classificação mais comuns incluem:

- Algoritmos de árvore: Constroem uma estrutura de árvore para representar diferentes caminhos de decisão com base nas características dos dados. Cada nó da árvore representa um teste em uma característica e cada ramo representa o resultado do teste. São métodos intuitivos e interpretaveis. Temos como exemplos: Árvore de decisão, *Random Forest*, entre outros.

- Regressão Logística: É usada para modelar a relação entre uma variável binária dependente e um conjunto de variáveis independentes. É útil quando o objetivo é classificar instâncias em duas categorias. A regressão logística estende o conceito da regressão linear para lidar com problemas de classificação, tornando-a eficaz quando as relações entre as variáveis são aproximadamente lineares.

- *K-Nearest Neighbors* (KNN): É um dos algoritmos mais clássicos e fundamentais. Ele pode ser usado tanto para classificação quanto para regressão, dependendo do contexto e dos dados disponíveis. No caso da classificação, o KNN é usado para atribuir uma classe ou categoria a uma instância de dados com base na maioria das classes dos K vizinhos mais próximos. O valor de K é um hiperparâmetro

que determina o número de vizinhos considerados na votação para decidir a classe atribuída.

- *Gradient Boosting*: Outro exemplo que pode ser usado tanto para classificação quanto para regressão. É uma abordagem de conjunto que combina a previsão de vários modelos fracos (geralmente árvores de decisão simples) para formar um modelo mais robusto e preciso. Quando usado para classificação, é empregado para prever a classe à qual um exemplo pertence, minimizando uma função de perda específica para a classificação. A implementação mais conhecida de *Gradient Boosting* é o algoritmo *Gradient Boosted Machine* (GBM), mas há também variantes populares como o *Extreme Gradient Boosting* (XGBOOST), *LightGBM* e *CatBoost*, cada um com suas próprias otimizações e melhorias.

2.4.3 Algoritmos de Regressão

Um problema de regressão em Aprendizado de Máquina refere-se a um tipo de tarefa em que o objetivo é prever um valor numérico contínuo com base nas características ou atributos das instâncias de dados. Diferentemente dos problemas de classificação, onde o objetivo é atribuir rótulos ou categorias, em problemas de regressão busca-se estimar um valor numérico específico.

Alguns algoritmos de regressão comumente utilizados incluem:

- Regressão Linear: Modelo que estabelece uma relação linear entre as variáveis independentes e a variável dependente. A regressão linear busca encontrar a melhor linha reta que se ajusta aos dados.
- *K-Nearest Neighbors* (KNN): Como já mencionado anteriormente, pode ser usado tanto para classificação quanto para regressão. No caso da regressão, o KNN é utilizado para prever um valor numérico contínuo para uma instância de dados. A previsão é baseada nos valores de saída (ou target) dos K vizinhos mais próximos.
- *Gradient Boosting*: Como também já mencionado anteriormente, é uma técnica que combina vários modelos fracos para obter um modelo forte. Quando aplicado a problemas de regressão, a meta é prever valores contínuos. Neste caso, a técnica cria sucessivas iterações de modelos, onde cada novo modelo é ajustado aos resíduos do modelo anterior.

2.4.4 Algoritmos de Agrupamento/Clusterização

Um problema de agrupamento em Aprendizado de Máquina refere-se a uma tarefa em que o objetivo é identificar estruturas e padrões nos dados, agrupando instâncias similares em conjuntos/grupos chamados *clusters*. O agrupamento é uma técnica exploratória

que visa encontrar similaridades entre os dados, independentemente das classes ou rótulos pré-definidos. Vale ressaltar que para alguns algoritmos é o programador que escolhe o número desejado de grupos.

Alguns algoritmos de agrupamento comumente utilizados incluem:

- *K-means*: É um algoritmo iterativo que particiona os dados em k *clusters*, onde k é um valor predefinido. Ele atribui cada instância ao conjunto mais próximo com base na distância euclidiana.
- *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN): É um algoritmo de agrupamento baseado em densidade que agrupa instâncias com alta densidade e separa regiões de baixa densidade. Ele não requer especificar o número de *clusters* antecipadamente e é capaz de identificar conjuntos de diferentes formas e tamanhos.
- Hierarchical Clustering: É uma abordagem que cria uma hierarquia de conjuntos, onde cada instância inicialmente é considerada um *cluster* separado e, em seguida, eles são mesclados com base na proximidade até formar um único *cluster*.

2.5 Avaliação do Desempenho de Modelos de Classificação

Segundo (VUJOVIC, 2021) existem várias métricas utilizadas para avaliar o desempenho de um modelo de aprendizado de máquina, e a escolha das métricas adequadas depende do tipo de problema e dos objetivos específicos. Conforme os diferentes tipos de problemas abordados na Seção 2.4 chegamos à conclusão que o trabalho em questão de predição de detratores em dados de telefonia se trata de um problema de classificação, que é da categoria dos aprendizados supervisionados. Para entender, portanto, as métricas relevantes para este projeto, é necessário entender a matriz de confusão.

2.5.1 Matriz de Confusão

Embora a matriz de confusão seja amplamente utilizada na área de aprendizado de máquina e classificação, seu conceito básico está relacionado à teoria de testes diagnósticos e estatística, onde é conhecida como tabela de contingência. A ideia de representar os resultados de classificação em uma matriz de confusão é intuitiva e prática para entender o desempenho de um algoritmo de classificação.

Portanto, a matriz de confusão é a matriz quadrada em que se compara os verdadeiros valores de uma classificação com os valores preditos através de algum modelo. Sua diagonal é composta pelos acertos do modelo e os demais valores são os erros cometidos. O caso binário, o mais comum, é representado pela seguinte matriz:

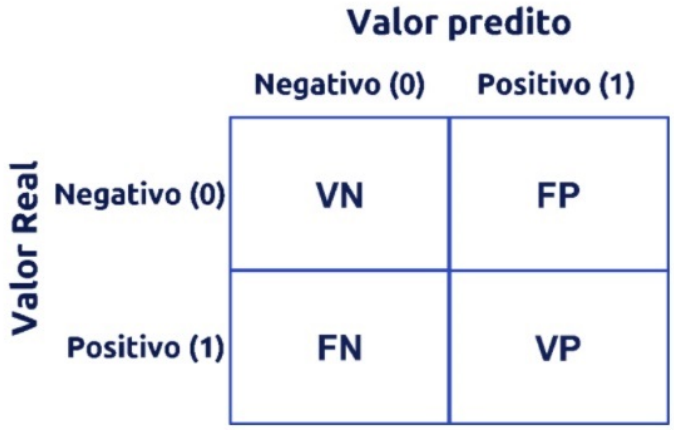

Figura 2 – Matriz de confusão explicada. Imagem disponível em: <https://www.flai.com.br/juscudilio/qual-a-melhor-metrica-para-avaliar-os-modelos-de-machine-learning/>

A matriz nos traz as informações das frequências dos acertos e erros do modelo. Ou seja, nos mostrará as frequências:

- Verdadeiro Negativo (VN): são as observações que o modelo previu como negativas e realmente eram negativas. Ou seja, o modelo classificou corretamente.
- Falso Negativo (FN): são as observações que o modelo identificou como negativas, mas eram positivas. Ou seja, o modelo classificou erroneamente.
- Verdadeiro Positivo (VP): são as observações que o modelo previu como positivas e realmente eram positivas. Ou seja, o modelo classificou corretamente.
- Falso Positivo (FP): são as observações que o modelo identificou como positivas, mas eram negativas. Ou seja, o modelo classificou erroneamente.

2.5.2 Métricas relevantes para este projeto

A seguir, estão listadas algumas das métricas que podem ser mais relevantes para o cenário dos dados de telefonia, com base na matriz de confusão:

- Acurácia (*Accuracy*): É a métrica mais básica e amplamente utilizada. Representa a proporção de predições corretas (VP + VN) em relação ao total de predições (VP + FP + VN + FN).

$$accuracy = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.5}$$

  É adequada quando as classes estão balanceadas e não há custos diferentes para os erros de classificação.

- Precisão (*Precision*): Indica a proporção de predições positivas corretas (VP) em relação ao total de predições positivas (VP + FP).

$$precision = \frac{VP}{VP + FP} \tag{2.6}$$

  É útil quando o foco é reduzir os falsos positivos.

- Revocação (*Recall*): Também conhecida como Sensibilidade ou Taxa de Verdadeiros Positivos, representa a proporção de instâncias positivas corretamente identificadas (VP) em relação ao total de instâncias realmente positivas (VP + FN).

$$recall = \frac{VP}{VP + FN} \tag{2.7}$$

  É relevante quando o objetivo é minimizar os falsos negativos.

- Medida F1 (*F1 Score*): É uma métrica que combina a precisão e a revocação em uma única medida. Ele é calculado como a média harmônica entre os dois e fornece uma medida balanceada do desempenho do modelo. O F1-score é particularmente útil quando há um desequilíbrio entre as classes e o objetivo é equilibrar a precisão e a revocação.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{2.8}$$

2.5.3 Outras métricas

- *Receiver Operating Characteristic* (ROC): A curva ROC é criada ao comparar duas razões: a Sensibilidade, também chamada de taxa de verdadeiros positivos, ou *recall* - Equação 2.7 - que é a proporção de verdadeiros positivos em relação ao total de positivos, e a Especificidade, também chamada de taxa de verdadeiros negativos, que é a proporção de verdadeiros negativos em relação ao total de negativos - Equação 2.9.

$$Especificidade = \frac{VN}{VN + FP} \tag{2.9}$$

  Isso é feito para diferentes valores de um limite de classificação. Em outras palavras, a curva ROC mostra como a capacidade de um modelo em identificar corretamente tanto os casos positivos quanto os negativos varia para diferentes ajustes do limite de decisão. Quanto mais próxima a curva estiver do canto superior esquerdo, como ilustrado na Figura 3, melhor o desempenho do modelo em equilibrar a sensibilidade e a especificidade para diferentes cenários de classificação.

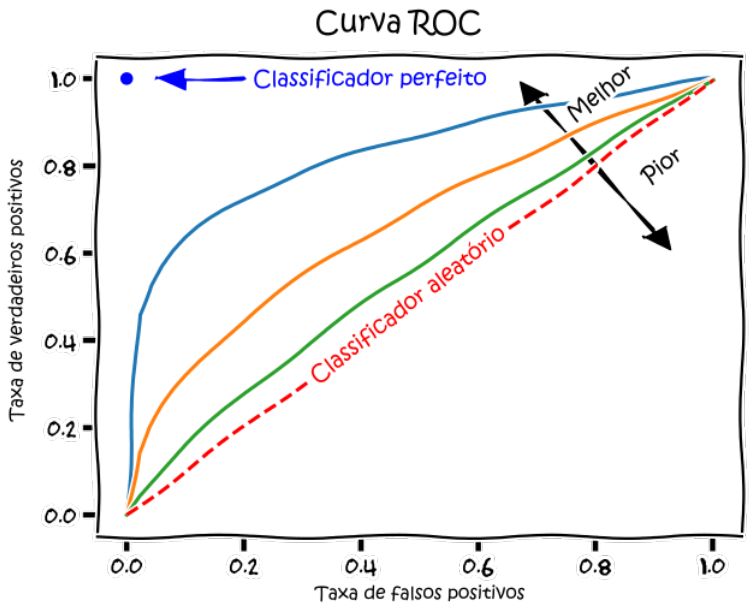

Figura 3 – Curva ROC. Imagem disponível em: <https://pt.wikipedia.org/wiki/Característica_de_Operação_do_Receptor>

- *Area Under the Curve* (AUC): Como o nome sugere, é a área abaixo da curva, onde essa curva é a ROC. A AUC representa a habilidade do modelo em discriminar entre as classes positiva e negativa. Ela não depende de um ponto de corte específico e fornece uma medida agregada do desempenho do modelo em toda a faixa de possíveis pontos de corte. A AUC é particularmente útil quando se deseja avaliar o poder de discriminação do modelo e comparar diferentes classificadores ou abordagens.

2.5.4 *Feature Importance*

A análise da importância das características (*Feature Importance*) é uma etapa importante da avaliação de um modelo, que visa identificar e avaliar a contribuição relativa de cada variável ou característica no processo de tomada de decisões do modelo. Isso permite não apenas a interpretabilidade do modelo, mas também fornece insights valiosos sobre quais características têm maior influência no resultado preditivo.

Existem diferentes métodos para calcular a importância de features, e a abordagem exata pode variar dependendo do algoritmo utilizado. Os próprios algoritmos de aprendizado de máquina possuem métodos para reportar as importâncias calculadas, bastando ao programador exibí-las da forma que achar mais conveniente. Isso pode ser feita com o uso de bibliotecas especializadas como a *SHapley Additive exPlanations* (SHAP). Entretanto, uma forma não rara de ser utilizada para visualização das *feature importances* é a simples disposição em um gráfico de barras.

2.6 Trabalhos Relacionados

Diversos trabalhos de predição utilizando dados de empresas de telecomunicação já foram realizados. Entretanto, o foco é, em sua grande maioria, na predição de *churns*, que são as saídas dos clientes, de fato, do serviço. Ou seja, trata da rotatividade de clientes. A maioria dos estudos aborda técnicas de aprendizado de máquina, mas alguns também utilizam redes neurais ou modelos híbridos. A seguir, alguns resultados de estudos.

O experimento conduzido por (HASSOUNA et al., 2015) aborda uma comparação entre variações dos algoritmos Árvore de Decisão e Regressão Logística na predição de *churns*. Foram usados dois conjuntos de dados - de 15.519 e 19.919 clientes contendo 17 variáveis - obtidos de uma operadora de telecomunicações móveis do Reino Unido, cujo nome não foi divulgado. Ambos os conjuntos de dados continham 50% de clientes que abandonaram e 50% que permaneceram na operadora, ou seja, as bases estavam balanceadas. O melhor resultado foi obtido para o modelo Árvore de Decisão C5, que obteve uma AUC de 76.3% e uma acurácia de 70.25%.

Já para (GAUR; DUBEY, 2018), quatro modelos foram empregados, foram eles: *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest* e *Gradient Boosting*. Os autores usaram a métrica AUC para medir o desempenho dos modelos. O modelo mais preciso foi o *Gradient Boosting* com valor de AUC de 84,57%. Os dados fornecidos são reais, porém não tiveram sua empresa responsável divulgada (ou a localidade da mesma), somente foi afirmado que a base contém 7.043 clientes e 21 variáveis. Também não foi informado se os dados estão ou não desbalanceados.

No estudo de (ULLAH et al., 2019) é proposto um modelo de predição de *churn* que utiliza técnicas de classificação e de agrupamento para identificar a rotatividade de clientes e apontar os fatores por trás disso. A base de dados utilizada para a predição de *churns* foi fornecida pela GSM, uma empresa de telecomunicações do sul da Ásia. Essa base contém 64.107 instâncias com 29 colunas, nos quais todas são numéricas. Há um desbalanceamento de 30% de clientes que abandonaram o serviço para 70% que foram retidos. O algoritmo *Random Forest* apresentou o melhor desempenho com 88.63% das instâncias classificadas corretamente.

Esse outro estudo, por (AHMAD; JAFAR; ALJOUMAA, 2019) teve como objetivo construir um sistema que prevê a rotatividade de clientes na empresa de telecomunicações SyriaTel, da Síria. Os dados utilizados nesta pesquisa contêm todas as informações dos clientes ao longo de nove meses antes da linha de base. O volume deste conjunto de dados é de cerca de 70 Terabytes em *Hadoop Distributed File System* (HDFS), e possui diversos formatos de dados estruturados, semiestruturados e não estruturados. Os dados também chegam muito rápido e precisam de um suporte adequado de uma plataforma de *big data* para lidar com isso. O conjunto de dados é agregado para extrair recursos para cada cliente.

Foram testados apenas algoritmos de árvore sob o argumento de que os dados não estavam balanceados - apenas 5% de *churn* - e tais algoritmos não são afetados por esse problema. Esses algoritmos foram: Árvore de Decisão, *Random Forest*, GBM e XGBOOST. Este último - que é um algoritmo baseado em árvore de decisão e utiliza uma estrutura de *Gradient Boosting* - obteve os melhores resultados em todas as medições. O valor de AUC para ele foi de 93,30%.

Por fim, neste outro estudo mais atual, quatro algoritmos de classificação - *Logistic Regression*, *Random Forest*, SVM e Árvore de Decisão - foram implementados e analisados por (COPACEANU, 2021). O objetivo dos modelos aqui é maximizar o *recall*, pois representa a capacidade do classificador de prever corretamente todos os casos verdadeiramente positivos. O melhor o valor de *recall* foi de 90.7%, obtido pelo modelo de Árvore de Decisão. Foi mencionado que a base de dados utilizada é uma base pública, sem mais informações de sua origem. Essa base possui 58 atributos e 51.047 clientes e contém desbalanceamento: 28.6% (14,257) de *churners* para 71.5% (35,519) de clientes retidos.

São menos numerosos os estudos que exploram o NPS nos dados de telefonia. Os dois estudos a seguir exemplificam as direções tomadas por diferentes autores.

Na abordagem de (MUSTAFA; LING; RAZAK, 2021), foram aplicadas técnicas de mineração de dados ao conjunto de dados NPS de uma empresa não divulgada de telecomunicações da Malásia em setembro de 2019 e setembro de 2020, analisando 7.776 registros com 30 campos para determinar quais variáveis eram significativas para o modelo de previsão de *churn*. O objetivo foi identificar os fatores por trás do *churn* de clientes além de propor uma estrutura de previsão do mesmo. Foram desenvolvidos e comparados modelos de previsão de *churn* usando Regressão Logística, Análise Discriminante Linear, KNN, *Classification and Regression Trees* (CART), *Gaussian Naïve Bayes* e SVM. O resultado obtido foi que a rotatividade de clientes é elevada para clientes com baixo NPS. O modelo CART apresentou a previsão de rotatividade mais precisa: 98% das instâncias classificadas corretamente. Vale ressaltar que a pesquisa está proibida de acessar informações pessoais de clientes de acordo com a política de proteção de dados da Malásia.

Finalmente, (MARKOULIDAKIS et al., 2020) utilizaram um conjunto de dados de pesquisas NPS realizadas no mercado de telecomunicações móveis da Grécia, sem mencionar uma empresa em específico. O conjunto de dados continha 9 atributos e 450 amostras, o que é muito pouco para treinar um modelo preditivo. Para solucionar, portanto, o problema da disponibilidade de dados, um gerador de dados é explorado para criar um conjunto de dados de pesquisa NPS com base nas principais estatísticas parâmetros da amostra original, como média, desvio padrão e matriz de correlação. Após esse enriquecimento da base de dados, obteve-se um total de 10.800 pesquisas como entrada para os modelos. A pesquisa de NPS realizada contava com a sua parte quantitativa e a qualitativa. A quantitativa se refere à nota dada ao serviço, enquanto a qualitativa é a

justificativa da nota, frequentemente tratados como atributos de experiência, ou *Customer Experience* (CX), como plano tarifário, qualidade do serviço, faturamento e experiências de pontos de contato, como *website*, *call center*, aplicativo móvel, etc. Os autores mencionam que o *dataset* está desbalanceado, mas não afirmam o quão desbalanceado. São aplicados modelos preditivos para classificação do NPS, ou seja, o modelo tenta prever se é Detrator, Passivo ou Promotor. Foram testados diversos modelos, são eles: Árvore de Decisão, KNN, SVM, *Random Forest*, Regressão Logística e ainda algumas redes neurais. Também foram feitos testes considerando diferentes informações passadas para os modelos. O que obteve melhor *f1-score* quando passadas todas as informações disponíveis foi o *Random Forest* com aproximadamente 79%.

Um resumo dos trabalhos citados, para melhor visualização e comparação, pode ser encontrado na Tabela 1. Em contraste à referência biliográfica apresentada, o presente trabalho pretende abordar o NPS com uma análise exploratória dos dados. Além disso, haverá uma predição, que não será, todavia, do *churn* - porque não temos essa informação na base de dados passada. Nem será mesmo, da classe NPS, mas, mais especificamente de clientes insatifeitos: os detratores. A relação entre insatisfação e abandono do serviço (detratores e *churn*, respectivamente), entretanto, é íntima, já que usuários insatisfeitos com o serviço prestado são os que possuem maior chance de mudarem de operadora.

Autor	**Base de dados**	**Metodologia**	**Melhor Algoritmo**	**Métrica Avaliadora**
(HASSOUNA et al., 2015)	Operadora do Reino Unido	Predição de *churn*	Árvore de Decisão C5	AUC = 76.3% e Acurácia = 70.25%
(GAUR; DUBEY, 2018)	Real; Operadora não informada	Predição de *churn*	*Gradient Boosting*	AUC = 84.57%
(ULLAH et al., 2019)	GSM (sul da Ásia)	Predição de *churn*	*Random Forest*	Acurácia = 88.63%
(AHMAD; JAFAR; ALJOUMAA, 2019)	SyriaTel (Síria)	Predição de *churn*	*Extreme Gradient Boosting*	AUC = 93.30%
(COPACEANU, 2021)	Real e pública; Operadora não informada	Predição de *churn*	Árvore de Decisão	*Recall* = 90.7%
(MUSTAFA; LING; RAZAK, 2021)	Operadora da Malásia	Predição de *churn* Análise do NPS	Classification and Regression Trees (CART)	Acurácia = 98%
(MARKOULIDAKIS et al., 2020)	Pequena amostra real + grande parte gerada	Predição da classe NPS	*Random Forest*	*F1-Score = 79%*
Presente Projeto	TIM (Brasil)	Predição de detratores Análise do NPS	*Random Forest*	*F1-Score = 61%*

Tabela 1 – Comparativo das referências bibliográficas e do presente projeto

3 Caracterização do *Net Promoter Score*

Este capítulo faz uma breve descrição do tratamento aplicado em cima dos dados fornecidos. Ademais, apresenta hipóteses levantadas na caracterização da (in)satisfação de clientes da operadora de telefonia móvel, assim como os resultados que sustentam essas hipóteses, cuja métrica avaliadora escolhida foi o *Net Promoter Score* (NPS).

Vale lembrar que as hipóteses são formuladas com base na visualização de gráficos e tabelas levantados dos dados disponíveis, que podem não refletir o cenário real quando enviesados, mesmo que sejam aplicadas diferentes técnicas para mitigar os efeitos de viés nos dados.

3.1 Metodologia

A estratégia empregada para caracterizar os clientes foi fazer o cálculo do NPS para cada *feature*, isoladamente, que permaneceu na tabela de dados após o tratamento da mesma. Todavia, apresentaremos apenas os resultados mais significativos, ou seja, aqueles que exibem discrepâncias notáveis em relação a determinadas categorias.

Os resultados estão apresentados em formato de tabela, composta pelos seguintes atributos: Percentual de promotores, de passivos e de detratores; Cálculo do NPS (%Promotores - %Detratores); e Percentual de distribuição da classe. Cada tabela de NPS apresentada está ordenada do melhor para o pior NPS (de forma decrescente), destaca-se em azul o melhor NPS dentre as categorias disponíveis e, em vermelho, o pior.

3.1.1 Acerca dos dados disponibilizados

A TIM coletou dados de seus clientes durante 6 (seis) meses - de julho a dezembro de 2022 - acerca da experiência de utilização dos serviços em dispositivos móveis. Esses dados contemplam um arquivo no formato *Comma Separated Values* (CSV) com 74 colunas x 46640 linhas, onde as colunas são as *features* e as linhas são os clientes entrevistados.

Para cada cliente que compõe a base de dados foi feita uma pesquisa de satisfação em duas etapas (perguntas). A primeira delas pediu para o cliente atribuir uma nota de zero a dez para o serviço da TIM e, a segunda, pediu a justificativa para a nota atribuída, por meio de opções prontas.

A TIM seguiu fielmente a divisão básica entre as categorias estabelecidas por Fred Reichheld - citada no Capítulo 2, Seção 2.1 - para classificação dos clientes em detratores, passivos ou promotores a partir da nota que os mesmos atribuíram ao serviço na pesquisa

de satisfação.

A tabela disponível no Apêndice A detalha todos os dados coletados pela empresa. Entretanto, um resumo do apêndice está disponível na Tabela 2. Ele contém as *features* essenciais para análise inicial juntamente com as que serão discutidas mais adiante por terem se destacado no processo de caracterização do NPS.

CAMPO	**DESCRIÇÃO**	**TIPO**
NPS_Class	Classe NPS (detrator, neutro ou promotor)	Categórico
USERVALUEQ2_VALUE	Atribuição à nota dada (sinal, estabilidade, etc)	Categórico
ANF	DDD do cliente	Numérico
term_fabr	Marca fabricante do dispositivo móvel	Categórico
tot_chamada	Total de chamadas	Numérico
originada_perc	Percentual de chamadas originadas	Numérico
qtde_celula	Quantidade de antenas utilizadas pelo usuário	Numérico
CUSTOMER_SEGMENT	Categoria do Plano (pré, controle ou pós)	Categórico
Qtd_CRM	Interações com CRM (duvidas, reclamações, etc)	Numérico
vol_Rede_Social	Consumo de dados em redes sociais	Numérico
vol_Navegacao	Consumo de dados em navegadores	Numérico
vol_MarketPlace	Consumo de dados na loja de aplicativos	Numérico
vol_Musica	Consumo de dados em música	Numérico

Tabela 2 – *Features* em destaque

3.1.2 Tratamento dos Dados

O tratamento dos dados para essa análise inicial consistiu na eliminação de colunas irrelevantes (com informações que não acrescentam na nossa análise como, por exemplo, data em que foi feito o contato com o cliente), colunas de valor único (preenchidas totalmente pelo mesmo valor) e colunas com mais de 20% de seus valores nulos, onde esse limite foi uma decisão de projeto.

Para as colunas com valores nulos que permaneceram, foi dado um tratamento individual para cada caso. Para as colunas categóricas, por exemplo, uma solução foi preencher com o texto 'não informado'. Já para as colunas numéricas, cada uma foi analisada separadamente e foram tomadas diferentes decisões como preencher com o valor mediano ou com o valor de maior ocorrência da coluna em questão.

Também foram analisados os valores discrepantes presentes em cada coluna numérica. Como não foram passadas as unidades de medida das variáveis, é difícil afirmar valores que são erros de medida, portanto, todos foram tratados como medidas corretas. Foi escolhida a abordagem do *Interquartile Range* (IQR), traduzido como Amplitude

Interquartil, para definir o que é *outlier* no *dataset*. Entretanto, com essa técnica, as *features* de volume apresentavam muitos *outliers*, por vezes com diferenças significativas em sua magnitude. Para tratar esse problema foi aplicada uma transformação logarítmica nos dados.

A técnica de transformação logarítmica é empregada para reduzir a variação nos dados. Isso é feito diminuindo os valores extremos proporcionalmente ao seu grau de distância da média - com a aplicação da função logarítmo (log) - o que os torna menos influentes na análise estatística do conjunto de dados. O resultado é a transformação de uma relação exponencial em uma relação linear. Dessa forma, não eliminamos valores discrepantes, mas mitigamos seu efeito por meio da redução de magnitude.

3.2 Análise Inicial

Após tratados os dados, uma análise inicial mostrou que, dentre todos os clientes entrevistados, 52% deram notas 9 ou 10 (Promotores), 18.5% deram notas 7 ou 8 (Passivos) e 29.5% deram notas entre 0 e 6 (Detratores) para o serviço prestado pela operadora de telefonia móvel TIM. A Figura 4 ilustra as distribuições das classes em valores absolutos e percentuais.

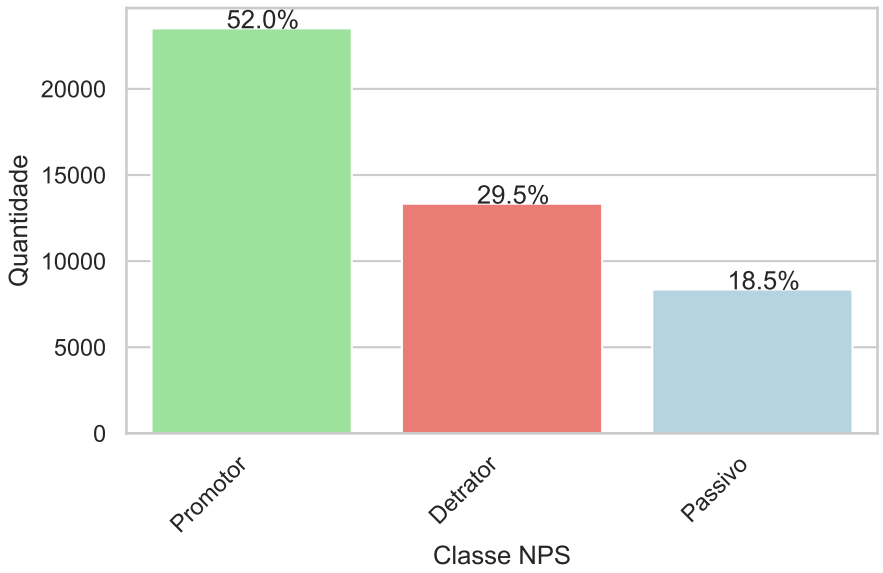

Figura 4 – Distribuição de Promotores, Passivos e Detratores (coluna "NPS_Class")

Além disso, ao exibir o gráfico de justificativas (segunda pergunta da pesquisa de satisfação), constatou-se que a maior queixa entre os que avaliaram o serviço com notas baixas é a cobertura de sinal. Entretanto, ao se exibir o mesmo para os promotores, a justificativa para os que avaliaram o serviço com notas altas também é a cobertura de sinal. Em ambos os casos liderando consideravelmente à frente das outras opções, como pode ser visto na Figura 5, referente aos detratores, e na Figura 6, referente aos promotores.

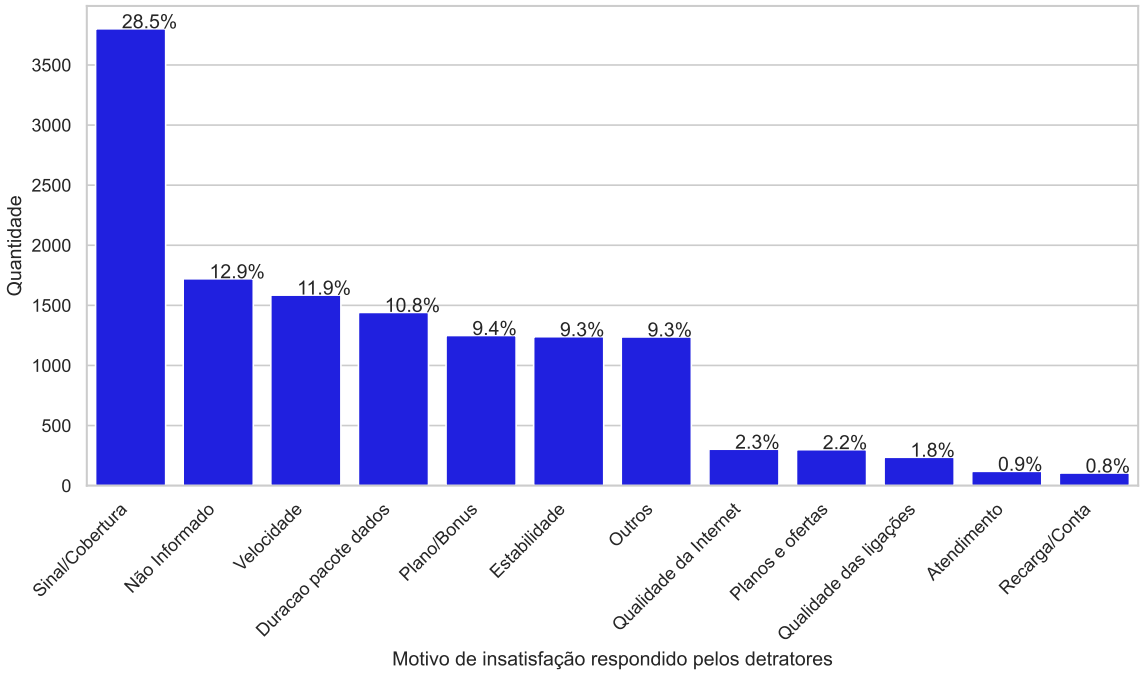

Figura 5 – Justificativa para a nota atribuída pelos detratores (coluna "USERVALUEQ2_VALUE")

.

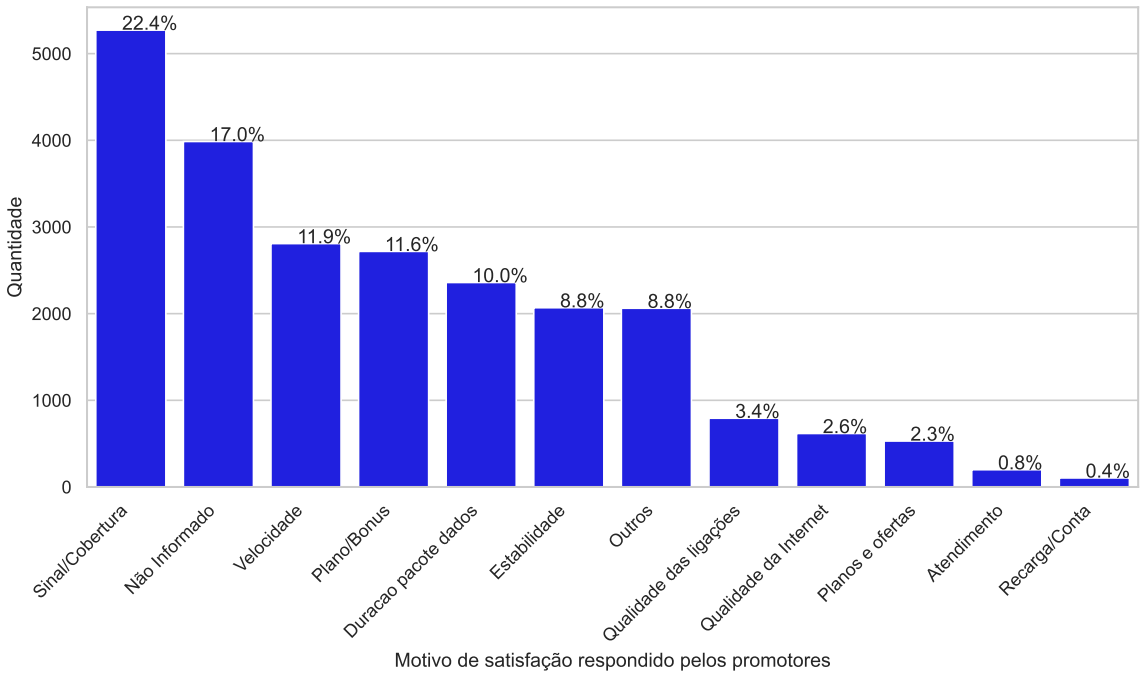

Figura 6 – Justificativa para a nota atribuída pelos promotores (coluna "USERVALUEQ2_VALUE")

A princípio não há grandes diferenças entre as respostas dos detratores e as repostas dos promotores, o que é contraintuitivo. É importante ponderar que diversos motivos podem contribuir para esse tipo de resultado - esses motivos são mais amplamente discutidos na

Seção 5.1. Todavia, nessa segunda pergunta da pesquisa, em especial, pode ser que por haverem muitas opções para o usuário ouvir e ponderar, o mesmo acaba desistindo de terminar de ouvir e, simplesmente vota na primeira (que coincidentemente, ou não, é a opção 'Sinal/Cobertura'), ou mesmo não vota, o que contribui para uma alta incidênia na classe 'Não Informado'.

Dessa forma, conclui-se que somente visualizando as respostas à pesquisa de satisfação não é possível se extrair muitos insights acerca do cenário de satisfação e insatisfação ou a respeito do perfil de promotores e detratores. Torna-se necessário uma análise mais minuciosa, que será feita com base no NPS.

3.3 Características Sociais e Demográficas

O primeiro cálculo do NPS foi feito para o estado e para a região dos clientes e está ilustrado na Tabela 3. Apesar de não haver campos nos dados referentes a essas informações (estado e região), é possível obtê-las com um agrupamento dos DDDs. Ao interpretar os resultados, surge a hipótese de que o *Net Promoter Score* pode estar intimamente relacionado com o Índice de Desenvolvimento Humano (IDH) de cada região, de forma que um maior IDH contribua para um menor NPS e vice versa. A lógica por trás disso seria que um dos pilares que compõem o IDH é a educação dos cidadãos, portanto, uma maior educação pode estar relacionada também com expectativas mais elevadas, maiores sensibilidades a problemas ou falhas, maior consciência das alternativas de mercado e maior exigência de modo geral.

Observe que as regiões Norte e Nordeste, que possuem um IDH mais discrepante das outras regiões, são as que possuem melhor NPS. De forma análoga, as regiões Centro-Oeste, Sul e Sudeste possuem maiores IDHs e piores NPSs quando comparadas as outras duas. Cada estado, porém, pode possuir suas particularidades, mas, no geral, o mesmo se repete.

Para sustentar a hipótese levantada de haver uma relação inversamente proporcional entre IDH e NPS, foi investigada a correlação linear entre as duas variáveis. O resultado obtido está ilustrado no mapa de calor da Figura 7 e foi obtido da correlação linear entre o NPS e o IDH para cada estado. O resultado de -0.51 indica uma significativa correlação inversa entre as duas variáveis.

Outra relação obtida que reforça ainda mais essa suspeita é o NPS por marca fabricante do dispositivo móvel de cada usuário, mostrado na Tabela 4.

A hipótese levantada a partir da análise dessa tabela é que, dentre as marcas dos aparelhos telefônicos, a Apple é a única marca que não possui diferentes opções de dispositivos variando o custo, eles trabalham com apenas um modelo de telefone que é atualizado anualmente e possui elevado custo no Brasil. Isso, atrelado à questão do poder

ESTADO			REGIÃO		
Sigla	**NPS**	**IDH**	**Nome**	**NPS médio**	**IDH médio**
AC	0.146	0,710	Norte	0.33	0.703
AM	0.421	0,700			
AP	0.329	0,688			
PA	0.333	0,690			
RO	0.131	0,700			
RR	0.286	0,699			
TO	0.041	0,731			
AL	0.239	0,684	Nordeste	0.294	0.702
BA	0.266	0,691			
CE	0.330	0,734			
MA	0.341	0,676			
PB	0.337	0,698			
PE	0.279	0,719			
PI	0.320	0,690			
RN	0.327	0,728			
SE	0.268	0,702			
DF	0.135	0.814	Centro-Oeste	0.169	0.757
GO	0.111	0.737			
MS	0.351	0.742			
MT	0.247	0,736			
PR	0.191	0,769	Sul	0.174	0.777
RS	0.151	0,771			
SC	0.156	0,792			
ES	0.209	0,771	Sudeste	0.215	0.778
MG	0.150	0,774			
RJ	0.265	0,762			
SP	0.209	0,806			

Tabela 3 – NPS e IDH por estado e por região político-administrativa. Os dados de IDH foram obtidos da pesquisa realizada pelo IPEA no ano de 2021 e estão disponíveis em: <http://www.atlasbrasil.org.br/>

aquisitivo estar intimamente relacionado com a educação corrobora para o quadro que já foi levantado onde os usuários Apple possivelmente possuem, em sua maioria, melhor situação econômica. Como consequência disso, são mais exigentes de uma forma geral.

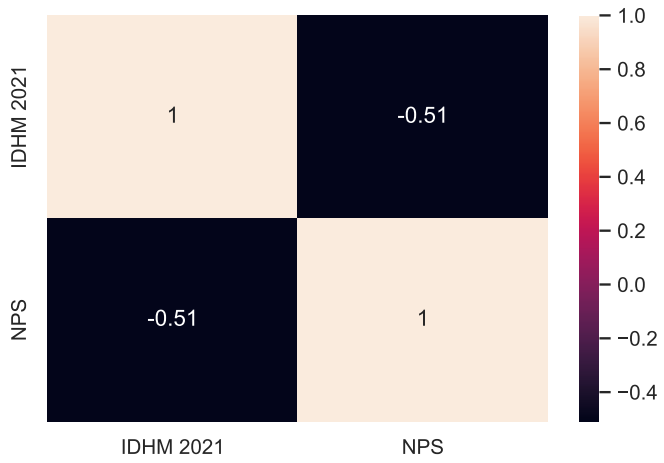

Figura 7 – Mapa de calor para correlação linear entre NPS e IDH de cada estado

Marca Fabricante	Detrator	Passivo	Promotor	NPS	Percentual
lg	25.31	17.18	57.51	32.2	7.59
outros	24.57	20.42	55.01	30.44	2.45
não informado	27.97	17.85	54.18	26.21	4.9
motorola	29.28	18.54	52.18	22.91	28.04
samsung	29.68	18.58	51.74	22.07	47.04
asus	32.93	19.51	47.56	14.63	1.27
xiaomi	34.18	17.24	48.58	14.39	4.35
apple	35.09	19.93	44.98	9.89	4.36

Tabela 4 – NPS por marca fabricante do dispositivo móvel (coluna "term_fabr")

3.4 Características de consumo de dados

As Tabelas 5, 6, 7 e 8 se referem ao cálculo do NPS de acordo com o consumo de dados em diferentes ambientes: redes sociais, navegadores, música ou loja de aplicativos.

Consumo de Dados	Detrator	Passivo	Promotor	NPS	Percentual
Baixo Consumo	26.9	19.09	54.01	27.11	33.83
Médio Consumo	29.62	18.41	51.96	22.34	35.74
Alto Consumo	31.18	18.52	50.3	19.13	13.86
Extremo Consumo	33.13	17.37	49.51	16.38	16.57

Tabela 5 – NPS por consumo de dados em redes sociais (coluna "vol_Rede_Social")

De maneira geral, quanto maior o consumo, maiores as chances de enfrentar problemas de sinal, estabilidade, entre outras questões que podem levar a uma experiência negativa com o serviço. Dessa forma, quanto maior o consumo, maiores as chances de o

Consumo de Dados	Detrator	Passivo	Promotor	NPS	Percentual
Baixo Consumo	27.79	18.12	54.09	26.3	78.84
Médio Consumo	34.39	19.94	45.67	11.27	8.91
Alto Consumo	35.02	20.66	44.32	9.3	4.85
Extremo Consumo	38.18	19.16	42.66	4.48	7.4

Tabela 6 – NPS por consumo de dados em navegadores (coluna "vol_Navegacao")

Consumo de Dados	Detrator	Passivo	Promotor	NPS	Percentual
Baixo Consumo	28.2	17.98	53.82	25.62	78.84
Alto Consumo	32.11	21.1	46.79	14.68	2.65
Médio Consumo	32.69	20.21	47.09	14.4	3.35
Extremo Consumo	35.1	20.25	44.65	9.54	15.16

Tabela 7 – NPS por consumo de dados em música (coluna "vol_Musica")

Consumo de Dados	Detrator	Passivo	Promotor	NPS	Percentual
Baixo Consumo	28.38	18.49	53.13	24.74	86.84
Médio Consumo	34.68	18.95	46.37	11.69	6.45
Alto Consumo	37.73	18.32	43.95	6.23	3.05
Extremo Consumo	39.94	17.7	42.36	2.42	3.65

Tabela 8 – NPS por consumo de dados na loja de aplicativos (coluna "vol_MarketPlace")

usuário se tornar um detrator. Observe que para todos os casos, o grupo de baixo consumo de dados apresenta melhor NPS, enquanto o grupo de extremo consumo de dados apresenta o pior NPS.

3.5 Características técnicas ou interativas

As Tabelas 9 e 11, analogamente às características de consumo de dados, refletem casos em que quanto maior a utilização e, principalmente, a dependência do serviço, maiores as chances do usuário ter experiências ruins e vir a se tornar um detrator.

Quando o usuário utiliza muito do recurso de ligações telefônicas, maiores as chances dele se deparar com quedas de sinal, oscilações, dentre outros problemas da rede que ocorrem no decorrer do dia, que podem levar à insatisfação. É interessante, entretanto, observar na Tabela 10 que um cenário onde o cliente recebe mais chamadas do que efetua colabora para um melhor NPS.

Da mesma forma, quanto maior a mobilidade do cliente, mais trocas de antenas irão ocorrer e maiores as possibilidades de se cair em uma antena que está com o sinal pior ou com maior instabilidade.

Consumo de Chamadas	Detrator	Passivo	Promotor	NPS	Percentual
Baixo Consumo	26.78	17.82	55.4	28.61	25.62
Médio Consumo	29.92	19.29	50.79	20.87	43.56
Alto Consumo	30.77	18.48	50.76	19.99	16.69
Extremo Consumo	31.61	17.22	51.17	19.56	14.13

Tabela 9 – NPS por consumo de chamadas telefônicas (coluna "tot_chamada")

Situação Chamadas	Detrator	Passivo	Promotor	NPS	Percentual
Mais recebidas	28.28	17.28	54.44	26.16	29.46
Mais originadas	29.85	18.99	51.16	21.31	57.99
Equilibrado	30.74	18.95	50.32	19.58	12.55

Tabela 10 – NPS por situação das chamadas (coluna "originada_perc")

Mobilidade	Detrator	Passivo	Promotor	NPS	Percentual
Baixa Mobilidade	25.21	16.78	58.01	32.8	28.3
Média Mobilidade	28.56	18.53	52.92	24.36	39.76
Alta Mobilidade	32.77	20.32	46.91	14.14	21.84
Extrema Mobilidade	38.16	19.12	42.72	4.56	10.1

Tabela 11 – NPS por mobilidade (coluna "qtde_celula")

Essas situações são comumente vivenciadas por pessoas que dependem do uso do telefone para trabalhar, seja viajando ou estando no mesmo local; postando nas redes sociais ou fazendo ligações. E, cada dia mais, essa realidade é experimentada pela população brasileira com o crescente movimento de digitalização, Segundo a (MEF, 2021), 66% dos brasileiros declararam que precisavam do smartphone no trabalho em 2021.

Agora tomemos na Tabela 12 o NPS por segmento/plano de contrato do cliente, que pode ser Pré-pago, Pós-pago ou Controle.

Plano	Detrator	Passivo	Promotor	NPS	Percentual
Pré-pago	25.77	15.55	58.68	32.92	34.84
Pós-pago	30.24	20.8	48.96	18.72	32.88
Controle	32.77	19.3	47.93	15.17	32.28

Tabela 12 – NPS por segmento de contrato/plano do cliente (coluna "CUSTOMER_SEGMENT")

Tal resultado nos leva a indagar por que pode haver uma diferença tão grande do modelo pré-pago para os outros dois modelos (pós-pago e controle). Assim, foram levantadas algumas hipóteses, mas, antes, é necessário entender como cada um dos planos funciona:

No plano pré-pago, os clientes pagam antecipadamente por um valor definido em créditos ou recargas. Não há faturas mensais nem contratos de longo prazo, assim, os usuários controlam seus gastos comprando recargas de crédito conforme necessário.

De forma antagônica, no plano pós-pago os clientes recebem uma fatura mensal com base no uso dos serviços no mês anterior. Normalmente, o plano pós-pago inclui um pacote de minutos, mensagens e dados, com a possibilidade de exceder esses limites e pagar taxas adicionais. Geralmente há um compromisso de contrato de longo prazo, geralmente de 12 a 24 meses, com multas por rescisão antecipada.

Já o plano controle pode ser considerado um equilíbrio entre os planos pré-pago e pós-pago. Nele, os clientes têm um limite de gastos mensais definido que não pode ser excedido. Como no plano pós-pago, geralmente há um contrato de longo prazo, mas a fatura é limitada ao valor pré-determinado. Esse plano oferece um controle maior sobre os gastos, evitando surpresas nas faturas, mas também limita a flexibilidade em relação ao pré-pago.

Dado o cenário observado onde os clientes dos planos pós-pago e controle possuem maior índice de insatisfação e, entendendo um pouco melhor sobre como funciona cada plano, algumas hipóteses podem ser levantadas, mas devem ser melhor investigadas pela operadora de telefonia. São elas: Os clientes em planos pós-pago e controle podem estar se surpreendendo com faturas mensais mais altas do que o esperado, especialmente se excederem os limites do plano ou não entenderem completamente a estrutura de preços; Os planos pós-pago e controle podem ter limites de dados para uso da internet. Se os clientes ultrapassarem esses limites, podem enfrentar redução de velocidade ou cobranças adicionais, o que pode levar à insatisfação.

Por fim, temos também, na Tabela 13, a comparação entre clientes que já abriram e clientes que nunca abriram chamados/reclamações na central de atendimento da telefonia

Situação Chamados	Detrator	Passivo	Promotor	NPS	Percentual
Nunca abriu chamado	28.68	18.57	52.75	24.07	93.69
Já abriu pelo menos 1 chamado	41.65	17.22	41.13	-0.53	6.31

Tabela 13 – NPS por interações com CRM: Reclamações, chamados, etc (coluna "Qtd_CRM")

Várias razões podem contribuir para o NPS (Net Promoter Score) ser significativamente inferior entre clientes de telefonia que já abriram pelo menos um chamado de reclamação em comparação com clientes que nunca abriram chamados. Algumas hipóteses levantadas são as seguintes: Clientes que já abriram chamados provavelmente tiveram experiências negativas ou problemas anteriores. Essas experiências podem já ter afetado negativamente a percepção deles em relação à operadora; Clientes que abriram chamados podem ter percebido problemas que persistem ou ocorrem com frequência; Se os problemas

relatados pelos clientes não foram resolvidos de maneira satisfatória, isso pode levar à frustração e à percepção de que a empresa não está comprometida em resolver efetivamente as questões.

3.6 Perfil do Cliente

Após minuciosas análises do NPS, a hipótese levantada acerca do perfil do cliente é a de que existem características em que sua ocorrência pode indicar uma maior probabilidade do usuário se tornar um detrator. São elas:

- Possuir um iPhone;
- Consumir muitos dados seja em redes sociais, navegando na internet, ouvindo música, navegando na loja de aplicativos ou outros cenários;
- Realizar muitas ligações telefônicas;
- Possuir alta mobilidade;
- Possuir plano controle ou pós-pago;
- Abrir chamados ou reclamações.

Do mesmo modo, também existem características em que sua ocorrência pode indicar maior probabilidade do usuário se tornar um promotor. São elas:

- Possuir um aparelho celular Android;
- Consumir poucos dados;
- Realizar poucas ligações telefônicas;
- Possuir baixa mobilidade;
- Possuir plano pré-pago;
- Não abrir chamados ou reclamações.

4 Predição de Detratores

Este capítulo classifica o tipo do problema em questão e mostra os resultados obtidos na construção de modelos de aprendizado de máquina para predição de clientes detratores em suas diferentes abordagens. Ao final, para o melhor modelo obtido é discutida a otimização feita e a importância das variáveis.

4.1 Metodologia

4.1.1 Ajustes Finais dos Dados

Além de todo o tratamento feito para a etapa de análise exploratória aprensentada no capítulo 3, foram necessários mais ajustes antes de se prosseguir para o treinamento de modelos de ML. Isso porque os modelos de aprendizado de máquina escolhidos operam com espaços vetoriais numéricos, portanto, não lidam com variáveis categóricas. Assim, foi necessário transformar todas as variáveis categóricas em numéricas e, para isso, foi utilizada a técnica de pré-processamento de dados conhecida como *One-Hot Encoding*. Essa técnica cria a quantidade necessária de variáveis binárias indicando a presença ou ausência de cada categoria.

4.1.2 Métricas para avaliação do desempenho

Por se tratar de um problema de classificação, as métricas utilizadas para avaliação do desempenho dos modelos foram a acurácia, a precisão, o *recall* e a *f1-score*. A métrica principal, que buscou ser maximizada nos modelos, foi a métrica *f1-score*, por ser considerada um equilíbrio entre precisão e *recall*.

4.1.3 Separação dos Dados em Treino e Teste

Foram separados 70% dos dados para treino e 30% para teste. Nos dados de treino foi implementada a técnica de validação cruzada, onde foram criados cinco grupos de validação. Dessa forma, as métricas apresentadas para cada modelo são médias obtidas de cada um desses cinco processos de treino, o que traz mais confiabilidade e credibilidade para os resultados. Enquanto isso, o conjunto de teste é o mesmo para todas as abordagens. Vale mencionar que o mesmo não foi "tocado"em momento algum, para não haver vazamento de dados.

4.2 Análise de Correlação

A primeira etapa foi verificar se existe correlação linear entre as *features* da base de dados. Como resultado, obteve-se que, apesar de haver correlações significativas entre algumas variáveis, não há correlação de nenhuma delas com a "NPS_Class_Detrator", que é a nossa coluna alvo, criada a partir do *One-Hot Encoding* para indicar se o cliente é detrator ou não (binária). É por isso que o problema em questão não é tão simples e faz-se necessário desenvolver modelos de ML que possam identificar padrões nos dados que vão além da linearidade.

4.3 Classificação do Problema

A predição de detratores nos dados de telefonia se trata de um problema de classificação, pois temos uma classe (Detratores) e queremos predizer se os clientes pertencem ou não a ela.

Problemas de classificação são problemas de aprendizado supervisionado já que, dada uma base de dados, temos uma coluna *target*, que é o nosso gabarito para medir o desempenho do modelo. Ou seja, temos um conjunto de dados rotulados em que cada entrada possui a saída correspondente e o objetivo dos modelos é aprender a fazer esse mapeamento de entradas e saídas.

4.4 Modelos Escolhidos

Após uma verificação inicial de alguns modelos em suas configurações *default*, foi decidido prosseguir para a otimização de hiperparâmetros somente com os modelos *Random Forest* e *Extreme Gradient Boosting*, por terem apresentado os melhores resultados iniciais. Após a otimização, o modelo *Random Forest* obteve as melhores métricas em todas as diferentes abordagens testadas, sendo assim, serão os resultados desse que serão discutidos ao longo do capítulo. Contudo, os resultados (matrizes de confusão) do *Extreme Gradient Boosting* também podem ser conferidos no Apêndice B.

Os ganhos da escolha desses modelos para tratar o problema em questão são: (1) São melhores em lidar com dados desbalanceados, o que é extremamente importante aqui já que, como visto na Seção 3.2 temos apenas 29.5% de detratores - que é o que queremos predizer - nos dados; (2) Permitem atribuir pesos diferentes às classes. Isso é útil quando você deseja dar mais importância à classe minoritária, tornando o modelo mais sensível a padrões nessa classe; (3) São mais robustos à *outliers* já que combinam a previsão de várias árvores para produzir uma saída final, o que também é de extrema importância aqui, pois, como foi discutido na Subseção 3.1.2, optamos por tratar todas as medições como corretas; (4) Não requerem normalização dos dados, o que traz mais simplicidade

para o projeto; (5) Nos trabalhos relacionados, foi observado que os algoritmos de árvore de fato se destacaram em vista dos outros ao serem treinados em dados de telefonia.

Por fim, destaca-se que os modelos abordados inicialmente, além dos já mecionados, foram: KNN, Regressão Logística e Árvore de Decisão.

4.5 Treino e Teste de Modelos de ML

Cinco abordagens diferentes de treino dos dados foram experimentadas em busca dos melhores resultados de *f1-score*. Essa métrica foi escolhida como principal já que o objetivo é predizer corretamente os detratores, mas não se sabe os custos dos diferentes erros cometidos pelo modelo. Assim, sendo a *f1-score* um equilíbrio entre precisão e *recall* - onde cada um pesa mais um determinado erro - é a melhor métrica para se levar em consideração nesse estudo. Entretanto, ao final, todas as métricas são dispostas em uma tabela para melhor visualização comparativa dos resultados.

4.5.1 Primeira abordagem: conjunto completo

Na Figura 8 está ilustrada a matriz de confusão obtida pelo modelo *Random Forest* no primeiro teste realizado, logo após a separação de treino e teste dos dados em sua completude. Na Tabela 14 está ilustrado o relatório de classificação contendo as métricas obtidas nessa primeira abordagem.

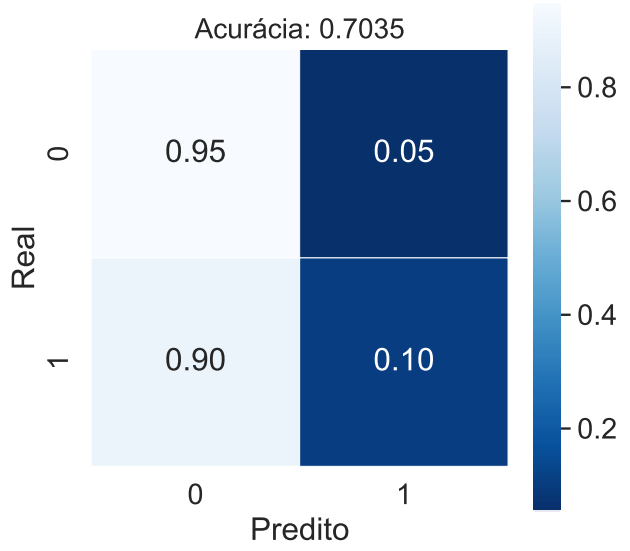

Figura 8 – Matriz de confusão obtida com *Random Forest* no treino com os dados completos

Observe que a acurácia do modelo pode ser considerada satisfatória, cerca de 70%. Entretanto, o que o modelo está fazendo, como se pode observar mais criteriosamente na distribuição da matriz de confusão, é dizer que nenhum cliente é detrator. Como há um

	precision	**recall**	**f1-score**	**support**
False	0.71	0.99	0.82	9539
True	0.51	0.03	0.06	4023
accuracy			0.70	
macro avg	0.61	0.51	0.44	13562
weighted avg	0.65	0.70	0.60	13562

Tabela 14 – *Classification report* obtido com *Random Forest* no treino com os dados completos

desbalanceamento nos dados - apenas 29.5% são detratores - já verificado na análise inicial (Seção 3.2), ao dizer que nenhum é detrator, realmente a acurácia vai ser em torno de 70%.

Ademais, independentemente das outras métricas - 65% de precisão, 70% de *recall* e 60% de *f1-score*. - o propósito aqui é saber quem são os detratores e, para isso, o modelo é incapaz de nos fornecer previsões, de forma que se torna inutilizável para o objetivo dado. Destaca-se que foi mostrada apenas a matriz de confusão para o modelo *Random Forest*, mas o mesmo resultado de concentração do resultado se repetiu para todos os outros modelos.

4.5.2 Segunda abordagem: extremos

Para prosseguir foi criada a hipótese de que os dados estão pouco claros para os modelos pelo seguinte motivo: Um cliente que dá, por exemplo, uma nota sete para o serviço, pode ter uma opinião pessoal de que essa nota é boa, mas outro cliente que dá a mesma nota para o serviço, pode ter uma opinião pessoal de que essa nota é ruim. Assim, esses clientes terão perfis e comportamentos diferentes tendo dado a mesma nota, o que confunde o modelo.

Uma alternativa pensada para evitar esse desentendimento é treinar o modelo apenas com os clientes que deram nota zero ou dez, pois, dessa forma, estaremos tratando as amostras mais distantes possíveis e teremos uma divisão mais clara de perfis de clientes. Posteriormente o modelo é testado em um conjunto de dados com amostras de todas as notas.

Os resultados obtidos aqui mudaram completamente em comparação à primeira abordagem. O modelo parou de classificar todos os casos como da mesma classe (promotores) e foi obtida a matriz de confusão ilustrada na Figura 9 e o relatório de classificação ilustrado na Tabela 15. As métricas obtidas para essa matriz foram as seguintes: 58% de acurácia, 64% de precisão, 58% de *recall* e 60% de *f1-score*.

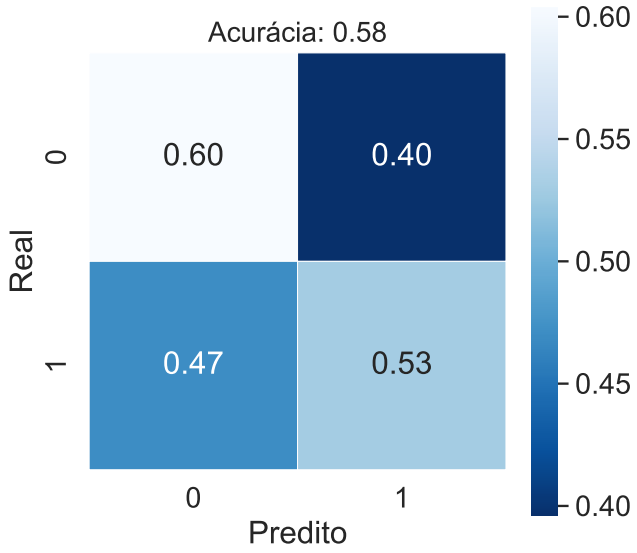

Figura 9 – Matriz de confusão obtida com *Random Forest* no treino somente com as notas zero e dez

	precision	**recall**	**f1-score**	**support**
False	0.76	0.61	0.68	9605
True	0.36	0.53	0.43	3957
accuracy			0.58	
macro avg	0.56	0.57	0.55	13562
weighted avg	0.64	0.58	0.60	13562

Tabela 15 – *Classification report* obtido com *Random Forest* no treino somente com as notas zero e dez

4.5.3 Terceira abordagem: extremos estendidos

Como tentativa de melhorar o modelo obtido com o treino nas notas zero e dez, por meio do fornecimento de um maior número de amostras, outra abordagem experimentada foi o treino com as notas zero, um, nove e dez, de forma que ainda temos extremos de perfis e um equilibrado número de amostras.

Os resultados obtidos estão exibidos na matriz de confusão da Figura 10 e no relatório de classificação da Tabela 16. As métricas obtidas foram as seguintes: 57% de acurácia, 63% de precisão, 57% de *recall* e 59% de *f1-score*.

Apesar do número mais significativo de amostras, os resultados de teste tiveram uma leve piora. A partir de então, tentativas de otimização nessa linha de raciocínio foram deixadas de lado.

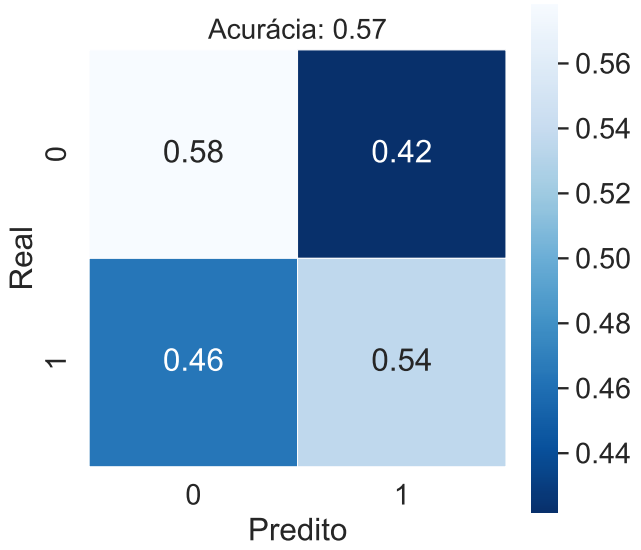

Figura 10 – Matriz de confusão obtida com *Random Forest* no treino somente com as notas zero, um, nove e dez

	precision	**recall**	**f1-score**	**support**
False	0.76	0.57	0.65	9605
True	0.35	0.55	0.43	3957
accuracy			0.57	
macro avg	0.55	0.56	0.54	13562
weighted avg	0.64	0.57	0.58	13562

Tabela 16 – *Classification report* obtido com *Random Forest* no treino somente com as notas zero, um, nove e dez

4.5.4 Quarta abordagem: amostras equilibradas

Outra alternativa pensada, em uma diferente linha, para impedir que o modelo preveja todos os clientes como da mesma classe é equilibrar as amostras dados de treino, mas mantendo casos de todas as notas. Assim, foram divididas amostras iguais de detratores e não detratores nos dados de treino, independente da nota dada.

Os resultados obtidos estão ilustrados na Figura 11 e na Tabela 17. Esses foram muito próximos dos resultados de treino nos extremos de notas zero e dez. O que difere aqui é que o modelo teve uma quantidade signitivamente maior de amostras para treino.

As métricas obtidas foram as seguintes: 59% de acurácia, 64% de precisão, 59% de *recall* e 61% de *f1-score*. Apesar de ter um resultado levemente superior nas métricas de acurácia e *recall* se comparado ao treino em cima das notas zero e dez, vale a discussão de qual modelo apresentaria melhores resultados caso treinados com o mesmo número de

amostras.

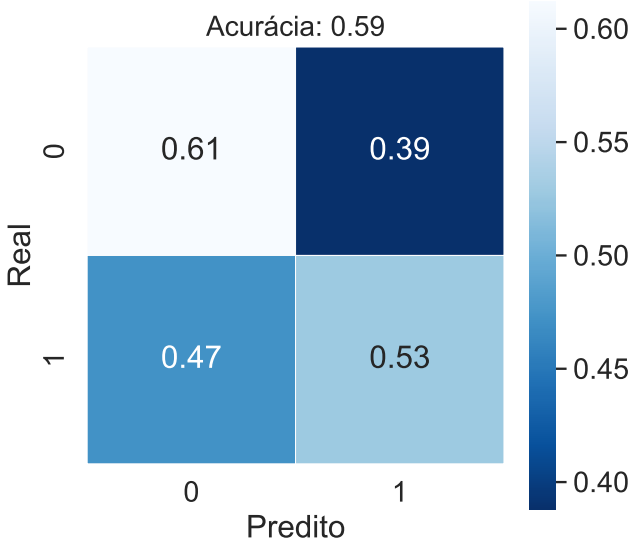

Figura 11 – Matriz de confusão obtida com *Random Forest* no treino com amostras iguais de detratores e não detratores

	precision	**recall**	**f1-score**	**support**
False	0.76	0.61	0.68	9605
True	0.36	0.53	0.43	3957
accuracy			0.59	
macro avg	0.56	0.57	0.55	13562
weighted avg	0.64	0.59	0.61	13562

Tabela 17 – *Classification Report* obtido com *Random Forest* no treino com amostras iguais de detratores e não detratores

4.5.5 Quinta abordagem: três classes

Por fim, a última abordagem experimentada na tentativa de se obter melhores resultados foi uma predição de três classes diferentes. Essas classes foram divididas da seguinte maneira:

- Classe 0: clientes certamente insatisfeitos - que deram notas 0, 1, 2 ou 3.
- Classe 1: clientes confusos para o modelo - que deram notas 4, 5, 6, 7 e 8.
- Classe 2: clientes certamente satisfeitos - que deram notas 9 e 10.

Apesar de não seguirem a divisão de Reichheld para detratores, passivos e promotores, identificar essas notas extremamente baixas já seria de grande serventia caso o modelo o consiga fazer com uma certa precisão.

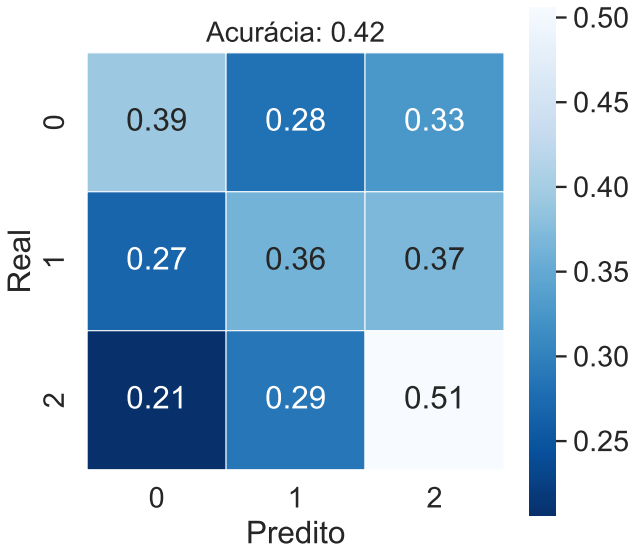

Figura 12 – Matriz de confusão obtida com *Random Forest* no treino com divisão de 3 classes

	precision	**recall**	**f1-score**	**support**
Classe 0	0.46	0.39	0.42	2368
Classe 1	0.38	0.36	0.37	2282
Classe 2	0.43	0.51	0.46	2370
accuracy			0.58	
macro avg	0.42	0.42	0.42	7020
weighted avg	0.42	0.42	0.42	7020

Tabela 18 – *Classification report* obtida com *Random Forest* no treino com divisão de 3 classes

Os resultados estão ilustrados na matriz de confusão da Figura 12 e no relatório de classificação da Tabela 18. É possível observar que o modelo se tornou melhor em identificar os clientes satisfeitos, mas não os insatisfeitos ou confusos, que eram o nosso maior objetivo. Com essa piora significativa em predizer os detratores, que são o foco principal, eventuais tentativas de otimização nessa linha de raciocínio também foram descartadas.

4.5.6 Comparativo das abordagens

A Tabela 19 traz, portanto, um comparativo das métricas para as diferentes abordagens de treino dos dados, que foram obtidas para o modelo *Random Forest*.

Abordagem	**Acurácia(%)**	**Precisão(%)**	***Recall*(%)**	***F1-Score*(%)**
Conjunto completo	70	65	70	60
Extremos	58	64	58	60
Extremos estendidos	57	63	57	59
Amostras equilibradas	59	64	59	61
Três classes	40	41	40	40

Tabela 19 – Comparativo das métricas de avaliação obtidas para cada abordagem de treino com o modelo *Random Forest*

Vale lembrar que a abordagem do conjunto completo, por mais que tenha obtido boas métricas, foi descartada por não inferir detratores, somente promotores, o que foge do objetivo dado. Já para as demais, a métrica que foi levada em consideração na escolha do melhor modelo foi a *F1-Score*.

É observado, entretanto, por meio dos *classification reports*, que para ambos os modelos das abordagens "Extremos", "Extremos estendidos"e "Amostras equilibradas"temos um *f1-score* comum de 43% para previsão da classe dos detratores. A diferença está, portanto, na classe dos não detratores, que o modelo de amostras equilibradas se mostrou melhor em predizer. Assim, o modelo com as amostras equilibradas foi o que obteve o melhor desempenho geral, logo após foi a abordagem dos extremos, seguida de uma leve piora do modelo quando esses extremos foram estendidos. Por fim, com o pior desempenho, temos a abordagem das três classes.

4.6 Otimização de Hiperparâmetros

A otimização por hiperparâmetros, como já comentado anteriormente, foi feita para os modelos *Random Forest* e *XGBoost*. Mas, como em todos os casos de teste o *Random Forest* apresentou melhor performance, será discutida apenas a otimização feita para esse modelo.

Os parâmetros otimizados estão listados a seguir, bem como a explicação do que se trata cada um, a faixa de valores fornecida para experimentação e o motivo para terem sido empregados no problema em questão:

"max_depth" : Profundidade máxima de cada árvore. Limitá-lo pode ajudar a prevenir *overfitting*, que é o extremo ajuste do modelo aos dados de treinamento, perdendo a

capacidade de generalização para novos dados. Também permite resultar em modelos mais simples, levando a um treinamento mais rápido e a uma menor demanda computacional. Isso é particularmente relevante em conjuntos de dados grandes. A faixa de valores fornecida para teste foi de 4 a 8.

"n_estimators" : Número de árvores do modelo. Aumentar esse parâmetro, especialmente em uma base grande de dados, pode permitir ganhos substanciais de desempenho. Entretanto, é importante encontrar um equilíbrio, pois aumentar indefinidamente o número de árvores pode resultar em custos computacionais mais altos sem ganhos significativos. A faixa de valores fornecida para teste foi de 100 a 250 *estimators*.

"min_samples_leaf" : É o número mínimo de amostras em cada folha. Ajustar esse parâmetro pode ajudar a evitar que a árvore crie folhas que são muito específicas para pequenos grupos de clientes, o que poderia ser um sinal de *overfitting*. A faixa de valores fornecida para teste foi de 2 a 15 amostras em cada folha.

"max_features" : Número máximo de variáveis a serem consideradas em cada árvore. Este hiperparâmetro é uma forma de regularização que influencia a diversidade entre as árvores do Random Forest. Restringir o número de features utilizadas em cada árvore aumenta a diversidade entre as mesmas. Isso pode melhorar a capacidade do modelo de generalização para diferentes padrões presentes nos dados, além de tornar o treinamento do modelo mais eficiente computacionalmente, já que se trata de um conjunto de dados com muitas colunas. A faixa de valores fornecida para teste foi de 2 a 10 *features*.

"subsample" : Dita a fração dos dados de treinamento que é usada para ajustar cada árvore individual. Em outras palavras, ele controla a amostragem aleatória dos dados antes de construir cada árvore no *Random Forest*. Controlar a fração dos dados usada para ajustar cada árvore pode ajudar na melhor generalização do modelo e na melhor eficiência computacional. A faixa de valores fornecida para teste foi de 0.7 a 0.9 (70% a 90% dos dados).

Como resultado, o melhor modelo obtido possui uma profundidade máxima de 7 árvores, um máximo de 8 *features* por árvore, um mínimo de 10 amostras em cada folha para 188 árvores e 90% dos dados.

4.7 *Feature Importance*

O modelo que apresentou as melhores métricas, que foi o *Random Forest* treinado com dados balanceados das duas classes (detratores e não detratores), teve o *top* 10 de *features* ilustrado na Figura 13.

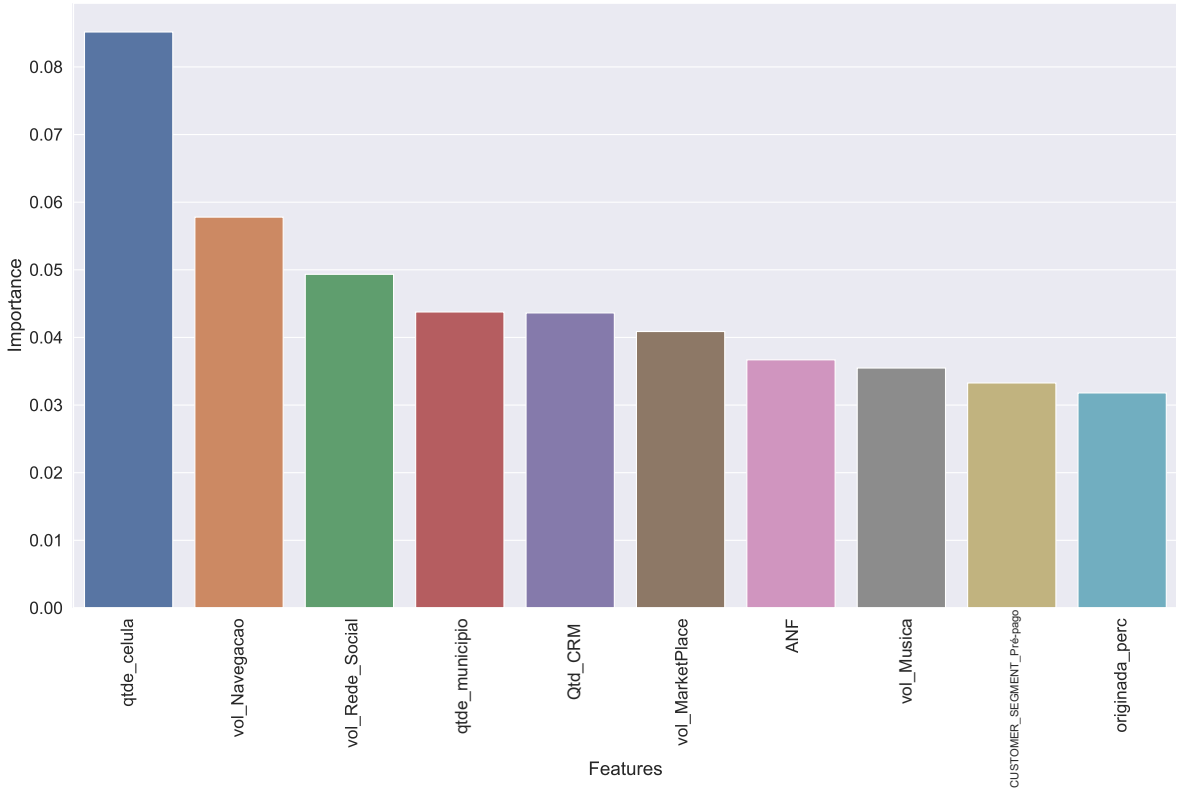

Figura 13 – Importância das variáveis para o modelo *Random Forest* treinado com dados balanceados de detratores e não detratores

Todas as *features* já haviam sido analisadas (com exceção da "qtde_municipio") - por meio do cálculo do NPS - e destacadas no Capítulo 3. Elas tiveram, portanto, sua importância comprovada pelo modelo. A razão para "qtde_municipio"não ter sido analisada é porque seu significado está fortemente correlacionado com a mobilidade (coluna "qtde_celula").

No capítulo 3, é possível observar que, a mobilidade do cliente ("qtde_celula"), dentre todas as características apresentadas nas seções 3.3, 3.4 e 3.5, foi a que apresentou maior queda do NPS entre o melhor NPS obtido, para baixa mobilidade, e o pior NPS obtido, para extrema mobilidade. Assim, faz muito sentido que o modelo tenha colocado essa *feature* com a maior importância, ou o maior peso, na hora de tomar suas decisões.

Em contrapartida, a análise de usuários Apple e não-Apple, que parecia também pesar na decisão, não apareceu entre as dez *features* mais importantes do modelo. Essa descoberta ressalta a importância do paralelismo entre uma abordagem analítica abrangente e o treino e avaliação dos modelos de predição. Esse trabalho em conjunto permite obter insights mais precisos e confiáveis sobre os fatores que impactam a satisfação e a lealdade dos clientes.

É importante também deixar claro que, devido a fatores de aleatoriedade dos modelos de aprendizado de máquina, a cada vez que eles são reinicializados, essas *features*

cuja importância está mais equilibrada podem variar de ordem. A disparidade, entretanto, da mobilidade como variável mais importante é evidente e bastante robusta.

Vale lembrar que, como já mencionado na Seção 2.5.4, o cálculo de *feature importance* varia de acordo com o algoritmo utilizado. Para algoritmos baseados em árvores, como o *Random Forest*, a importância de uma *feature* é calculada considerando a redução média na impureza que a a mesma proporciona nas várias decisões da árvore. Quanto maior a redução na impureza, mais importante a variável é considerada. O processo geral é o seguinte: (1) Para cada árvore na floresta é calculada a redução na impureza causada por cada feature em cada nó da árvore; (2) Essas reduções são agregadas para cada *feature* em todas as árvores da floresta; (3) Os resultados são normalizados para se obter uma pontuação de importância que reflete a importância relativa das variáveis no conjunto de dados.

5 Considerações Finais

5.1 Desafios e limitações nos dados de telefonia

Apesar do tratamento aplicado nos dados, alguns desafios não puderam ser eliminados por se tratarem de limitações instrínsecas à uma base real de dados. Um desafio recorrente é a questão do desbalanceamento, que ocorre quando uma classe ou categoria de interesse em um conjunto de dados é representada por um número significativamente menor ou maior de observações em comparação com outras classes. Isso pode ter um impacto substancial na qualidade e eficácia das análises e modelos, pois podem causar viés na análise ou modelagem e acurácia enganosa.

No presente projeto, para mitigar, em alguns momentos, o efeito do desbalanceamento uma técnica recorrentemente utilizada foi pegar amostras aleatórias de tamanho fixo para cada classe ou categoria em análise. O efeito dessa ação é satisfatório, mas impede que seja utilizada a base de dados completa.

Outra limitação é a veracidade das informações. A pesquisa do NPS feita, com base em duas perguntas já mencionadas anteriormente, pode apresentar resultados enganosos por vários motivos, dentre os principais: Os clientes podem ser influenciados por normas sociais e não quererem parecer críticos ou negativos. Isso pode levar a uma tendência de dar classificações mais altas do que o que realmente sentem; Se os clientes não perceberem benefícios em fornecer respostas sinceras ou críticas, podem optar por responder de forma superficial ou simplesmente não responder à pesquisa; Alguns clientes podem não entender completamente a pergunta ou podem interpretá-la de maneira diferente, o que pode resultar em respostas imprecisas; Momentos de irritabilidade do cliente podem levá-lo a dar uma nota muito baixa que não representa a real avaliação dele no geral e sim somente naquele momento isolado.

Não somente é preocupante a veracidade da informação em relação à pesquisa feita aos clientes, mas também nos dados obtidos por meio de medições como, por exemplo, os dados de volume. Não possuir as unidades de medida para cada variável numérica é um obstáculo que atrapalha a inferir com maior precisão se certas medições seriam *outliers* ou não.

Por fim, outra situação presente que prejudica bastante a parte de análise e caracterização, é a ausência de uma documentação completa. Por mais que modelos preditivos sejam indiferentes à documentação, não saber o significado de algumas *features* pode prejudicar a compreensão das *feature importances* e do cenário como um todo, pois não podemos trabalhar com suposições dos dados. Essa documentação incompleta é, infeliz-

mente, muito comum no mundo real, especialmente em grandes empresas, onde há maior número e rotatividade de colaboradores e, caso um deles não documente detalhadamente seu código, o outro pode não entender, ou mesmo entender errôneamente o trabalho.

5.2 Conclusão

Foi desenvolvido nesse projeto, uma análise do NPS para cada variável presente na tabela de dados fornecida pela operadora de telefonia móvel TIM. A partir disso, foram selecionados os resultados mais importantes para serem trazidos para discussão e, assim, descrever brevemente as tendências dos perfis opostos de clientes (detratores e promotores).

Ademais, foram treinados diferentes algoritmos de aprendizado de máquina diante de variadas separações dos dados. O modelo que apresentou o melhor resultado em todas as situações foi o *Random Forest*. O melhor resultado obtido desse modelo, foi para o treino nos dados com as classes balanceadas, onde foram mantidos casos de cada uma das notas atribuídas pelos clientes. A taxa *f1-score* no teste foi de 61% na situação descrita.

Espera-se que a TIM possa utilizar os insights fornecidos ao longo do projeto e o recurso de predição para o sucesso e crescimento sustentável de seu negócio em um cenário competitivo e dinâmico entre as operadoras de telefonia móvel. Além disso, cabe à telefonia despender esforços para verificar as hipóteses levantadas.

Apesar dos bons resultados já obtidos pelos modelos de aprendizado de máquina, que comprovam a existência de sinais nos dados de um comportamento comum entre grupos de clientes, existem pontos ainda podem ser aperfeiçoados. As sugestões de melhorias para trabalhos futuros estão descritas na próxima seção.

5.3 Trabalhos futuros

Ambas as etapas de carcterização e de predição apresentam grande potencial para melhorias, algumas delas foram identificadas durante o processo de desenvolvimento das mesmas e ficam como sugestões para trabalhos futuros.

Como sugestão para trabalhos futuros na caracterização da (in)satisfação de clientes de uma operadora de telefonia móvel, temos:

- Além do *Net Promoter Score* (NPS), existem várias outras métricas que podem ser usadas para entender o grau de satisfação dos clientes em relação ao serviço prestado por uma empresa. Duas métricas, em especial, podem ser muito interessantes nesse caso, são elas o *Customer Satisfaction Score* (CSAT) e o *Retention Rate* (Taxa de Retenção). Elas não foram abordadas nesse projeto por fugirem do escopo proposto, mas não deixam de ser interessantes de serem avaliadas para continuações do trabalho.

Agora como sugestões para trabalhos futuros na predição de detratores de uma operadora de telefonia móvel, temos:

- Levamos em consideração a métrica *f1-score* como principal, pois não foram atribuídos custos para os erros cometidos. Assim, pode ser solicitado à operadora de telefonia TIM que estime esses custos. Com os custos aproximados seria possível dar mais relevância à outra métrica de avaliação do desempenho como a precisão ou o *recall*, de forma a direcionar melhor o investimento da empresa.

- Seria interessante, também, obter mais dados dos clientes detratores para se ter um maior balanceamento das classes. Além disso, seria possível disponibilizar um maior número de casos dos mesmos para treino, proporcionando um melhor entendimento do alvo pelo modelo.

- Outra situação que pode ser testada nos modelos de predição, considerando que há um desbalanceamento considerável entre as classes é a atribuição de pesos. Isso se mostra vantajoso quando há o interesse em atribuir maior relevância à classe menos frequente. Foi mencionado que isso é possível de ser feito com os modelos de árvore.

- Como já mencionado anteriormente, é interessante verificar qual modelo se destaca mais quando treinados com a mesma quantidade de dados, dentre o modelo treinado somente com as notas 0 e 10 e o treinado equilibrando-se as classes, mas mantendo todas as notas nos dados.

- Pode ser implementada uma etapa a mais no projeto, que seria a busca por *outliers* multivariados. Esses são os dados que podem, perfeitamente, existirem sozinhos, mas que não fazem sentido de existirem em conjunto. Como por exemplo: se o volume de dados gasto por um cliente na rede social é de 30 e seu volume total de dados gasto é de 20, então esse é um caso de *outlier* multivariado e deveria ser eliminado, pois há uma inconsistência nos dados.

Por fim, para se obter ganhos tanto na etapa de análise quanto na etapa de predição, é interessante estudar como pode ser feita a junção dessa base de dados utilizada nesse projeto com outras bases que a TIM possui sobre seus clientes. Em especial, bases que tenham alguma informação de *churn*, que é a saída de fato do serviço e não somente a insatisfação gerada por ele.

Referências

AHMAD, A. K.; JAFAR, A.; ALJOUMAA, K. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 2019. Citado 2 vezes nas páginas 28 e 31.

BISHOP, C. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado na página 21.

COPACEANU, A.-M. Churn prediction in telecommunications sector using machine learning. *Database Systems Journal*, v. 12, 2021. Citado 2 vezes nas páginas 29 e 31.

DOE, J.; SMITH, J. Understanding the pearson correlation coefficient: A brief guide. *Journal of Statistical Analysis*, 2022. Citado na página 20.

GAUR, A.; DUBEY, R. Predicting customer churn prediction in telecom sector using various machine learning techniques. *International Conference on Advanced Computation and Telecommunication (ICACAT) and IEEE*, 2018. Citado 2 vezes nas páginas 28 e 31.

HASSOUNA, M. et al. Customer churn in mobile markets: A comparison of techniques. *Canadian Center of Science and Education*, v. 8, n. 6, p. 224–237, 2015. Citado 2 vezes nas páginas 28 e 31.

MARKOULIDAKIS, I. et al. A machine learning based classification method for customer experience survey analysis. 2020. Citado 2 vezes nas páginas 29 e 31.

MEF. *Mobile Communication For Employees Converging Private and Professional Lives*. 2021. Mobile Ecosystem Forum. Citado na página 40.

MUSTAFA, N.; LING, L. S.; RAZAK, S. F. A. Customer churn prediction for telecommunication industry: A malaysian case study. 2021. Citado 2 vezes nas páginas 29 e 31.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 18.

RAJU, V. N. G. et al. Study the influence of normalization/transformation process on the accuracy of supervised classification. *IEEE*, 2020. Citado na página 18.

REICHHELD, F. The one number you need to grow. *Harvard Business Review*, v. 81, n. 12, p. 46–55, 2003. Citado na página 15.

ULLAH, I. et al. A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE*, v. 7, p. 60134–60149, 2019. Citado 2 vezes nas páginas 28 e 31.

VUJOVIC, Classification model evaluation metrics. v. 12, 2021. Citado na página 24.

Apêndice A

Observações: Nem todas as *features* da base de dados foram documentadas pela TIM, de forma que o significado de algumas permaneceu oculto durante todo o projeto. Essas ainda puderam contribuir com o treinamento de modelos de predição, entretanto, para a análise e caracterização precisaram ser descartadas. As *features* cujos detalhes não se têm informações à respeito foram sinalizadas com '-' no campo 'DESCRIÇÃO'. Ademais, com tantas informações, a tabela construída a seguir precisou ser dividida em três partes para melhor visualização.

CAMPO	DESCRIÇÃO	TIPO	OBSERVAÇÕES
ref_month	Mês de contato com o cliente	Data	
CONTACT_DT	Data da pesquisa	Data	
ANF	DDD do Cliente	Numérico	
SURVEY	Tipo da pesquisa (Dados, app, etc)	Categórico	
reference_month	Mês referente à pesquisa	Data	
CUSTOMER_SEGMENT	Categoria do Plano (Pré, controle, pós)	Categórico	
USERVALUEQ1_LABEL	Nota dada pelo usuário de 0 à 10	Numérico	
USERVALUEQ2_VALUE	Atribuição à nota dada (Dados, fatura, app, etc)	Categórico	
NPS Class	Classificação em Detrator(0-6), Neutro(7-8) e Promotor(9-10)	Categórico	
vol_Rede_Social	Consumo de volume de dados referente à categoria	Numérico	
vol_Video		Numérico	
vol_Comunicacao		Numérico	
vol_Loja_de_Apps		Numérico	
vol_Musica		Numérico	
vol_Google		Numérico	
vol_Navegador		Numérico	
vol_InternetBank		Numérico	
vol_Transporte		Numérico	
vol_Jogos		Numérico	
vol_Marketplace		Numérico	
vol_TIM		Numérico	
vol_Alimentacao		Numérico	
vol_Ecommerce		Numérico	
vol_Viagem		Numérico	
desc_exp_flg	-	Binário	
desc_ativo_survey_flg	-	Binário	
qtd_acesso_com_trafego_fatura	Quantidade de usuários vinculados à conta	Numérico	mês atual / mês anterior
vlr_fatura	Valor da fatura no mês	Numérico	
flg_debito_automatico	Debito automatico ativo	Binário	
price_up_12m	Indica se o valor do plano subiu nos últimos 12 meses	Binário	
price_up_3m	Indica se o valor do plano subiu nos últimos 3 meses	Binário	
var_fatura_3m	Indica se houve variação da fatura nos últimos 3 meses	Binário	
trocas_de_plano_12m	Indica se houve trocas de plano nos últimos 12 meses	Binário	

CAMPO	DESCRIÇÃO	TIPO	OBSERVAÇÕES
qvar_vlr_fatura_3m	-	Binário	
ativ_ult_plan_dias	Tempo do ultimo contrato ativo	Numérico	
valor_recarga_sucesso	Somatório do valor das recargas	Numérico	
qtd_recarga_sucesso	Somatório da quantidade de recargas	Numérico	
perc_recarga_sucesso	Percentual de tentativas com sucesso	Numérico	
efic_recarga	-	Numérico	
FLAG_EXISTENCIA_R	-	Numérico	
FLAG_PARÂMETRO_R	-	Numérico	
FLAG_ELEGIBILIDADE_R	-	Numérico	
FLAG_VOLTE_R	-	Numérico	
imsi_LTE	Chip habilitado para LTE	Binário	
qtde_celula	Quantidade de células utilizadas pelo usuário	Numérico	
tot_chamada	Total de chamadas	Numérico	
tot_duracao_chamada	Total de minutos	Numérico	
cham_csp_dif_41_perc	-	Numérico	
tot_sms_env_receb	Total de SMS enviado e recebidos	Numérico	
tot_sms_receb	-	Numérico	
originada_perc	Percentual de chamdas originadas	Numérico	tot_chamada_originada / tot_chamada
eficiencia_Voz	-	Numérico	
tot_volume	Tráfego de Dados UL e DL	Numérico	
vol_zero_rating_perc	-	Numérico	
vol_gratuito	-	Numérico	
vol_4g_perc	Tráfego de dados 4G	Numérico	tot_volume_4g/tot_volume
vol_5g_perc	Tráfego de dados 5G	Numérico	tot_volume_5g/tot_volume
vol_ul_perc	-	Numérico	
eficiencia_pdp	Eficiência de conexões de dados	Numérico	
qtde_municipio	Quantidade de municípios	Numérico	Mobilidade já prevista no atributo qtde_celula
VoLTE_perc	-	Numérico	
term_fabr	Marca fabricante do dispositivo móvel	Categórico	
term_class	-	Categórico	
device_tech	Tecnologia do Aparelho (2G, 3G, 4G, 5G)	Categórico	
DISP_STATUS_p	Disponibilidade das células mais utilizadas	Numérico	
ACD_STATUS_p	Acessibilidade das células mais utilizadas	Numérico	
OCUP_STATUS_p	Ocupação das células mais utilizadas	Numérico	

CAMPO	DESCRIÇÃO	TIPO	OBSERVAÇÕES
Qtd_CRM	Total de interações com CRM (duvidas, reclamações, etc)	Numérico	
tent_acesso_app_meutim	-	Numérico	
eff_acesso_app_meu_tim	Eficiência de acesso ao App Meu TIM	Numérico	
dias_franquia_zerada	Dias com saldo zerado (Somente pré e controle)	Numérico	
dias_traf_red	Qtde de dias com tráfego reduzido (somente pós-pago)	Numérico	
qtd_pct_adic_dados	-	Numérico	
qtd_pct_adic_redes_soc	-	Numérico	

Apêndice B

Resultados dos modelos de *Extreme Gradient Boosting* otimizados.

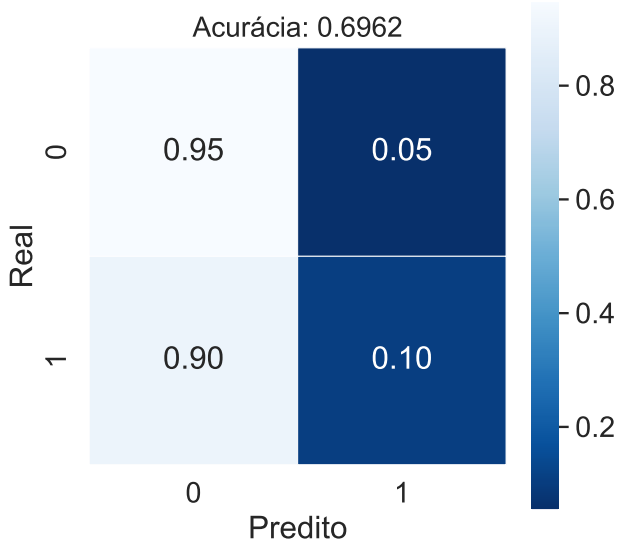

Figura 14 – Matriz de confusão obtida com *Extreme Gradient Boosting* no treino com os dados completos

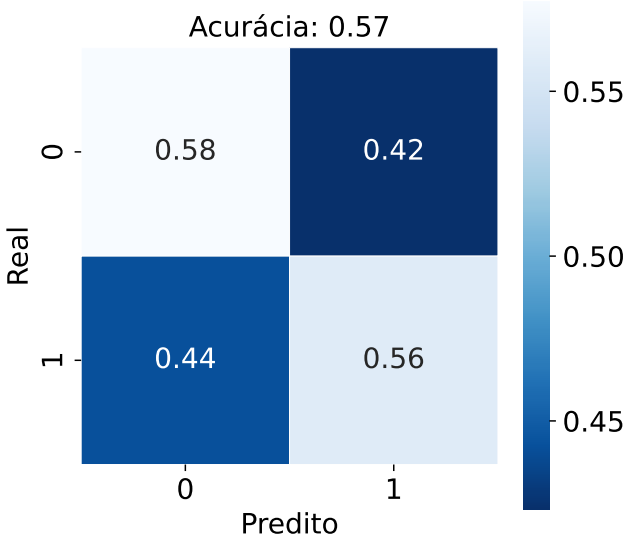

Figura 15 – Matriz de confusão obtida com *Extreme Gradient Boosting* no treino somente com as notas zero e dez

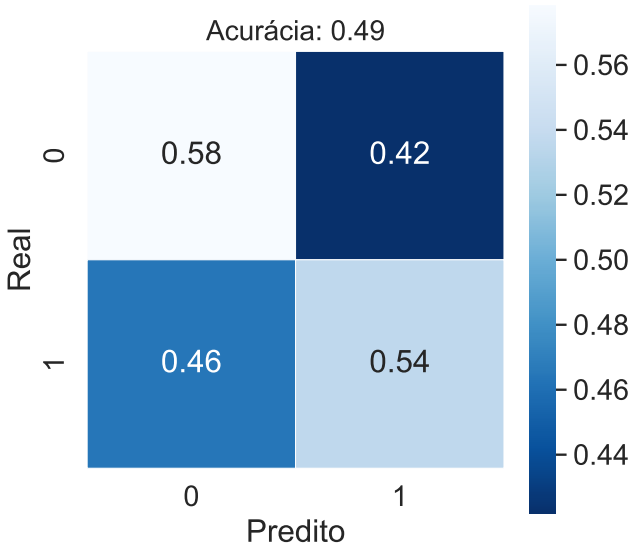

Figura 16 – Matriz de confusão obtida com *Extreme Gradient Boosting* no treino somente com as notas zero, um, nove e dez

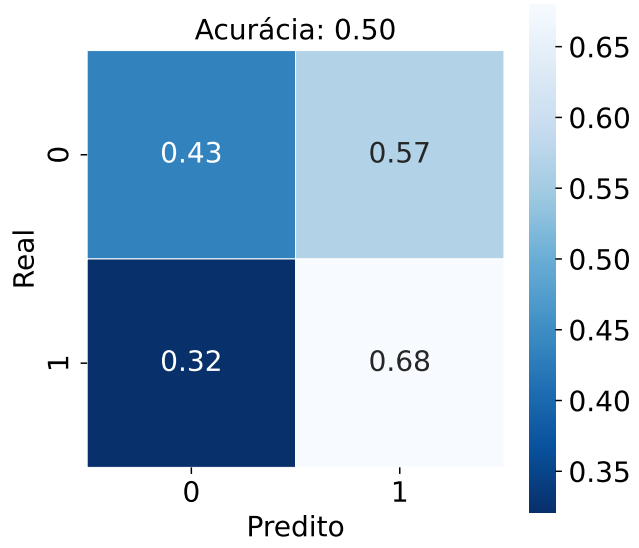

Figura 17 – Matriz de confusão obtida com *Extreme Gradient Boosting* no treino com amostras iguais de detratores e não detratores

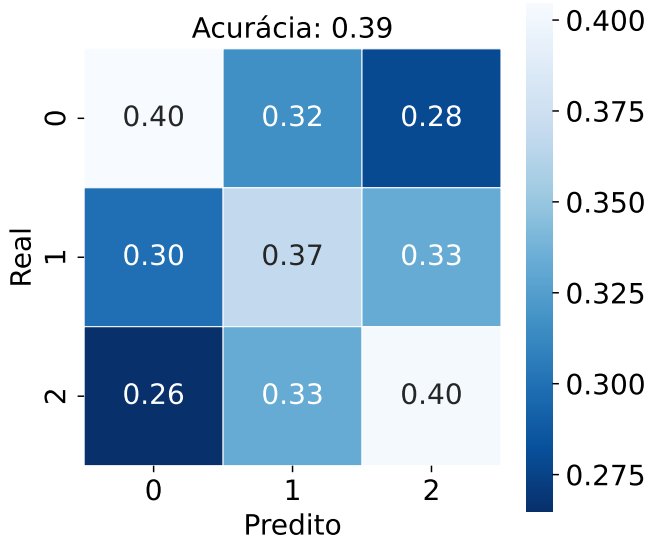

Figura 18 – Matriz de confusão obtida com *Extreme Gradient Boosting* no treino com divisão de 3 classes

Abordagem	**Acurácia(%)**	**Precisão(%)**	***Recall*(%)**	***F1-Score*(%)**
Conjunto completo	70	64	70	60
Extremos	57	64	57	59
Extremos estendidos	49	63	49	51
Amostras equilibradas	50	64	50	52
Três classes	39	39	39	39

Tabela 20 – Comparativo das métricas de avaliação obtidas para cada abordagem de treino com o modelo *Extreme Gradient Boosting*