PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação Lato Sensu em Ciência de Dados e Big Data

José Nunes da Silva Junior

ANÁLISE CHURN RATE TELECOM

Belo Horizonte

2023

José Nunes da Silva Junior

ANÁLISE DE CHURN RATE TELECOM

Trabalho de Conclusão de Curso apresentado

ao Curso de Especialização em Ciência de

Dados e Big Data como requisito parcial à

obtenção do título de especialista.

Belo Horizonte

2023

**SUMÁRIO**

Sumário

[1 INTRODUÇÃO 4](#_Toc147655298)

[1.1 Contextualização 4](#_Toc147655299)

[1.2 O problema proposto 4](#_Toc147655300)

[2 Coleta de Dados 5](#_Toc147655301)

[3 Tratamento de Dados 6](#_Toc147655302)

# INTRODUÇÃO

## Contextualização

Nos últimos anos, o setor de telecomunicações tem observado um enorme avanço tecnológico, avanço este, que possibilitou novas oportunidades de negócios para as empresas. A facilidade de contratar determinados serviços com poucos cliques também é encontrada ao se tentar cancelar este mesmo serviço. Dada a crescente competição no mercado bem como o aumento da demanda dos consumidores por serviços mais personalizados e acessíveis, um elemento vem se tornando o foco das atenções: o *churn rate*.

O “*churn rate*”, ou taxa de rotatividade de clientes, é uma métrica crucial para empresas no ramo de telecomunicações, diferentemente de um modelo de compra e venda onde se obtêm o lucro no momento em que o processo se concretiza, o modelo de negócios de telecomunicações se caracterizam por um investimento inicial feito pela empresa, seja em infraestrutura, equipamentos e/ou instalação na residência do cliente. Nesse modelo de negócio, o cliente só passa a dar lucro meses ou até mesmo anos depois de contratar o serviço.

Diante disso, tornou-se fundamental para as empresas o foco nesse indicador, uma vez que a retenção de clientes é crucial para o sucesso e a sustentabilidade dos negócios.

## O problema proposto

Devido as características de investimento inicial alto e lucro futuro, a presente analise priorizara no RECALL, uma vez que devemos buscar minimizar o maximo possivel de falsos negativos, onde o modelo determina que o cliente não vai cancelar porem o cliente cancela

# Coleta de Dados

O conjunto de dados *Telecom Customer Churn Prediction*, objeto de análise deste projeto pode ser encontrado através do sítio da *Maven Analytics*, uma plataforma de desafios relacionados a analises de dados.

Figura 1 - Fonte da base de dados que será utilizada.



Fonte: .

Na Figura 1 é possível observar os nomes e propriedades dos arquivos que serão utilizados. O conjunto em questão consiste em 3 (três) *datasets* que possuem informações e finalidades distintas (Figura 2): *telecom\_customer\_churn*, que contêm a base principal dos dados que serão analisados neste artigo; *telecom\_zipcode\_population*, que possui dados adicionais relacionados a população por código postal e; *telecom\_data\_dictionary*, que como o próprio nome sugere, trata-se de um dicionário que contém as informações do que cada coluna representa em nossa base principal.

Figura 2 - *Datasets* que serão utilizados



Fonte: Autoria própria.

# Tratamento de Dados

Para a execução do projeto, foi utilizada a linguagem de programação interpretada Python, na sua versão 3.9.13, além do *Jupyter Notebook*, ambiente de desenvolvimento interativo baseado na Web

Nesta etapa do projeto, será utilizada majoritariamente a biblioteca *Pandas* para manejo dos dados bem como de eventuais ajustes que possam ser necessários para deixar a base de dados melhor estruturada, e com isso, buscar melhores interpretações dos algoritmos de *Machine Learning*.

Figura 3 - Importação do *pandas*



Fonte: Autoria própria

Foi utilizado para a leitura dos dados, uma função própria do *pandas*, a *pd.read\_csv* (Figura 4), através dela, é possível ler um arquivo *csv* e já transforma-lo em um *dataframe* que nada mais é do que uma estrutura de dados composta por duas dimensões: linhas e colunas, similar a uma planilha.

Figura 4 - Leitura do arquivo com o pandas



Fonte: Autoria própria

Em seguida, por meio do comando *display* é possível visualizar o *dataframe* criado (Figura 5)

Figura 5 - Visualização do dataframe



Fonte: Autoria própria

Um detalhe importante sobre a função display é que ela por padrão, exibe no máximo 20 colunas, para que ela exiba todas as colunas e assim conseguirmos ter uma visão geral do nosso dataframe iremos realizar uma configuração através do comando *pd.set\_option* e alterando o parâmetro *display.max\_columns* como mostrado na Figura 6.

Figura 6 - Ajuste no maximo de colunas da função display



Fonte: Autoria própria

Dando continuidade agora realizaremos a leitura e criação do dataframe de população por código postal (), esses dados serão agregados ao nosso dataframe principal com o objetivo de enriquecer nossa base de dados, afim de encontrar algum padrão com a quantidade de pessoas de uma determinada região, o código usado para realizar essa junção está representado na Figura 8.

Figura 7 - Criando dataframe da população por código postal



Fonte: Autoria própria

Figura 8 - Junção dos dataframes



Fonte: Autoria própria

Agora temos o nosso *dataframe* consolidado e pronto para ser realizado os devidos ajustes

BIBLIOGRAFIA

O que são data frames? <https://www.databricks.com/br/glossary/what-are-dataframes>. Acessado 08/10/2023