

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

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

**RELATÓRIO TÉCNICO**

UM MODELO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO PARA PREVISÃO DE QUANTIDADE DE PÚBLICO NOS JOGOS DO CAMPEONATO BRASILEIRO DE FUTEBOL

Salomão Fernandes de Freitas Júnior

Belo Horizonte

2024

**SUMÁRIO**

[1. Introdução 3](#_Toc154172152)

[2. Entendimento do Domínio do Problema 4](#_Toc154172153)

[2.1. Contexto e Requisitos do Problema 4](#_Toc154172154)

[2.2. Definição da Estratégia a ser utilizada 6](#_Toc154172155)

[2.3. Escolha e Obtenção das Bases de Dados 6](#_Toc154172156)

[3. Preparação da Base de Dados 8](#_Toc154172157)

[3.1. Ambiente Tecnológico (????) 8](#_Toc154172158)

[3.1. Análise Exploratória dos Dados 8](#_Toc154172159)

[3.2. Seleção de Atributos – Redução de Dimensionalidade 8](#_Toc154172160)

[3.3. Análise de Valores Vazios (Nulos) 9](#_Toc154172161)

[3.4. Análise de Outliers 9](#_Toc154172162)

[3.5. Engenharia de Atributos (feature enginering) 9](#_Toc154172163)

[3.5.1 Criação de Novos Atributos 9](#_Toc154172164)

[3.5.2 Transformação de variáveis numéricas em categóricas 9](#_Toc154172165)

[3.5.3 Agrupamento de categorias (variáveis categóricas) 9](#_Toc154172166)

[3.5.4 Nova Seleção de Atributos 9](#_Toc154172167)

[3.6. Padronização de Variáveis 9](#_Toc154172168)

[3.7. Codificação de Variáveis Categóricas 9](#_Toc154172169)

[4. Escolha do Modelo 10](#_Toc154172170)

[5. Implantação da Solução em Ambiente de Produção 10](#_Toc154172171)

[6. Links 10](#_Toc154172172)

[REFERÊNCIAS 11](#_Toc154172173)

# 1. Introdução

Atualmente muitas das técnicas de Análise de Dados, Business Intelligence (BI) e Ciência de Dados, vêm sendo aplicadas para aprimorar processos decisórios em todos os aspectos dos esportes, uma área em franca ebulição denominada análise de dados esportivos (*sports analytics*) [1. Sharda, Ramesh].

Em nosso recente Trabalho de Conclusão da Pós-Graduação em Analytics e Business Intelligence [referência módulo A] [referência módulo B] [referência módulo C], concluído em novembro de 2022, pela PUC-Minas, já abordamos esse tema. O referido trabalho se constituiu de três módulos, os quais abordaram respectivamente, os seguintes objetos:

- Modelagem e Construção de um Data Warehouse do Campeonato Brasileiro da era de pontos corridos (2003 a 2021);

- Desenvolvimento da camada de apresentação da solução de BI, com um *Dashboard*, com três níveis de perspectivas interligadas entre si: Estratégico, Tático e Operacional, exibindo informações a partir dos dados do Data Warehouse;

- Construção de um modelo simples de aprendizado de máquina (machine learning) supervisionado, utilizando a técnica de Classificação, através de um algoritmo de Árvore de Decisão, com o objetivo de predizer o resultado de jogos (VITÓRIA, EMPATE, DERROTA), sob a perspectiva de um clube específico.

Neste atual projeto acadêmico, seguimos na mesma temática do nosso trabalho prévio, citado anteriormente. Durante o decorrer das disciplinas da Pós-graduação em Ciência de Dados e Big Data, foi possível obter um aprofundamento do aprendizado das técnicas envolvidas em todas as etapas de um projeto de ciência de dados, bem como dos principais algoritmos de aprendizado de máquina e dos aspectos envolvidos na sua utilização.

Dessa forma, apresentamos nesse trabalho um modelo de aprendizado máquina supervisionado, utilizando a técnica de Regressão, para predição de quantidade de público nos jogos do campeonato brasileiro de futebol.

O objetivo principal deste trabalho, foi a consolidação de conceitos, técnicas e abordagens tratados durante as disciplinas do curso de Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data, com o desenvolvimento de um projeto completo de ciência de dados, em todas as suas etapas.

A figura 1, a seguir, ilustra as etapas de um projeto de Ciência de Dados, as quais detalhamos nos próximos tópicos, dentro do contexto do projeto desenvolvido neste trabalho.

FIGURA: ETAPAS DE UM PROJETO DE CIÊNCIA DE DADOS (VER REFERÊNCIAS)

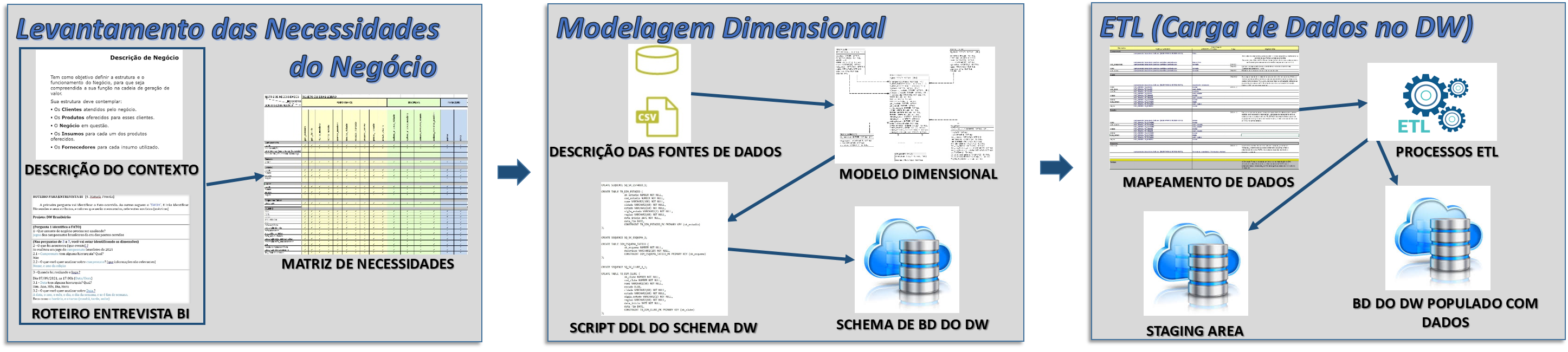


FIG. 1 – Etapas de Projeto de BI Abordadas no Módulo A do Relatório Técnico

# 2. Entendimento do Domínio do Problema

(Ver Disciplina Preparação de Dados para Machine Learnig)

(Ver citações nas referências)

Conforme descrito em [9.Zárate, Luis], nesta primeira etapa do projeto, o cientista de dados deve, juntamente com o especialista do negócio, entender e caracterizar o domínio do problema.

Devem ser bem entendidos o contexto e os requisitos do problema, identificadas as bases de dados disponíveis na organização ou mesmo externamente, identificando-se as características (atributos) relevantes que possam ser obtidos diretamente nas bases de dados disponíveis, ou que possam ser derivados destes.

Nos próximos tópicos descrevemos a etapa de Entendimento do Domínio do Problema no âmbito do nosso projeto.

## 2.1. Contexto e Requisitos do Problema

No nosso projeto, desejamos realizar a predição de quantidade de público nos jogos do campeonato brasileiro de futebol. Essa predição deverá ser realizada baseada em determinadas características que podem ser relevantes na determinação do interesse do público em um determinado jogo.

A predição do público esperado para um jogo, pode ser útil para determinação da logística envolvida na realização do mesmo, como confecção e venda de ingressos, quantidade de alimentos e outros itens a serem disponibilizados para venda no estádio, quantidade de força de segurança a ser disponibilizada dentro e fora do estádio, quantidade de efetivo e horário de funcionamento de transporte público, dentre outros aspectos. Também, o organizador do jogo poderá reforçar as estratégias de marketing, visando o maior incentivo da torcida naqueles jogos onde, pelas suas características, espera-se um baixo comparecimento de público.

Nosso escopo envolve somente o campeonato brasileiro de futebol, que caracteriza-se por ser um campeonato de “pontos corridos”, onde não existem jogos eliminatórios. O desempenho de um clube se define pela quantidade de pontos conquistados durante todo o campeonato.

Dessa forma, tentamos entender as características envolvidas em um jogo, que podem ter influência no interesse de uma torcida em comparecer ao mesmo, para prestigiar o time, podendo citar as seguintes:

- **Etapa do campeonato**: espera-se que as pessoas tenham um interesse menor no jogos iniciais, quando não há ainda uma definição clara de classificação;

- **Desempenho Geral de seu clube**: após passados os jogos iniciais, já se tem uma tendência de desempenho dos clubes, e um clube com bom desempenho naquele momento do campeonato, pode se mostrar postulante a ser campeão, ou a obter boas colocações, o que espera-se que gere incentivo em sua torcida a comparecer nos seus jogos locais;

- **Importância do clube adversário**: esta pode ser uma característica que motive uma torcida a comparecer a um jogo de seu clube na sua cidade. Clubes com maior tradição ou relevância, em visita a uma cidade, podem influenciar no interesse de uma torcida local a comparecer ao estádio;

- **Desempenho recente do seu clube**: também, mesmo que um clube não esteja em uma boa colocação no campeonato, espera-se que caso apresente um bom desempenho nos últimos jogos, isso gere um inventivo a sua torcida, pela tendência de crescimento. Da mesma forma, um clube bem colocado no campeonato, mas com desempenho ruim nos jogos mais recentes, pode causar desmotivação a sua torcida;

- **Dia da semana ou horário do jogo**: o dia de realização de um jogo (fim de semana ou dia útil) e o horário do mesmo (manhã, tarde, noite), podem interferir no interesse ou possibilidade de comparecimento de torcida;

- **Período do ano**: alguns meses podem apresentar tendência de maior ou menor comparecimento de torcida.

Além dos atributos citados, outras características mais julgamentais, como por exemplo o perfil de torcida, ou mesmo a condição climática, ou até o acontecimento de outros eventos na cidade, podem influenciar no comparecimento de público em um jogo.

**(VER EXPLICAÇÃO DO TERMO JULGAMENTAIS – DISCIPLINA DO CURSO)**

## 2.2. Definição da Estratégia a ser utilizada

Dado o entendimento inicial do problema e de seus requisitos, descritos no tópico anterior, foi possível definir a estratégia de solução do mesmo.

Assim, optamos por utilizar uma técnica de aprendizado de máquina supervisionado, já que queremos predizer o valor de um atributo (Quantidade de Público), levando em consideração dados históricos presentes em um conjunto de dados (*Dataset*).

Algoritmos de aprendizado supervisionado, basicamente, tentam modelar relacionamentos entre valores de atributos de entrada e suas saídas (atributos-alvo) correspondentes, a partir de dados de treinamento. Assim, tornam-se capazes de predizer valores de saída para novos dados de entrada. [2.Sarkar, Dipanjan]

Existem dois principais tipos de métodos de aprendizado supervisionado: **Classificação**, no qual o objetivo é predizer valores de saída para um atributo categórico (conhecidos como classes); e **Regressão**, que por outro lado, objetivam a estimação de valores de saída para atributos numéricos contínuos, conforme descrito em [2.Sarkar, Dipanjan] e [3.Guido, Sarah].

No nosso projeto, utilizamos a técnica de **Regressão**, uma vez que nosso atributo-alvo (Quantidade de Público) é um valor numérico, que embora seja representado por um número inteiro, devido à grandeza e variabilidade do mesmo, podemos tratá-lo como contínuo. Nossos atributos de entrada serão, a princípio, aqueles citados no tópico 2.1, conforme disponibilidade dos mesmos no conjunto de dados históricos (*Dataset*) que teremos disponível.

Nosso modelo deverá “aprender” com os dados históricos (os quais devem possuir registros que contenham tanto os atributos de entrada quanto o atributo-alvo com valores preenchidos), e deverá ser capaz de predizer o valor do atributo-alvo para novos dados de entrada.

## 2.3. Escolha e Obtenção das Bases de Dados

Após o entendimento geral do domínio do problema, já pudemos estabelecer uma estratégia para a solução. No entanto, precisamos verificar a disponibilidade das informações (atributos) definidas como requisitos da solução. Essas informações podem estar disponíveis nas bases de dados da própria organização para quem se desenvolve a solução, ou poderão ser obtidas externamente.

Para o nosso projeto, realizamos uma pesquisa geral na internet, bem como em sites mais específicos, conhecidos da comunidade de ciência de dados, destacando-se o *Kaggle* [<https://www.kaggle.com/datasets>] e o *Base Dos Dados* [<https://basedosdados.org/>].

Dentre os *datasets* encontrados, selecionamos aquele disponibilizado em

[<https://basedosdados.org/dataset/c861330e-bca2-474d-9073-bc70744a1b23?table=18835b0d-233e-4857-b454-1fa34a81b4fa>].

O *dataset* **brasileirao\_serie\_a.csv**, selecionado, possui a maioria dos atributos que definimos como requisito da nossa solução, e inclusive o nosso atributo-alvo (Quantidade de Público), com dados históricos do campeonato brasileiro de futebol da era de pontos corridos. O *dataset* disponibilizado de forma gratuita, contém dados de 29/03/2003 até 26/10/2023.

Para preenchimento dos dados faltantes (de 27/10/2023 a 7/12/2023), de modo a contemplar todos os jogos realizados até o momento da escrita deste trabalho, utilizamos as informações contidas no site [[https://www.srgoool.com.br](https://www.srgoool.com.br/)], tendo sido possível completar quase todas as informações de interesse, as quais concentramos em um segundo *dataset*, **brasileirao\_serie\_a\_dados\_complementares\_2023.csv**.

Dessa forma, para o nosso projeto, obtivemos dois *datasets*, os quais abrigam dados de todos os jogos do campeonato brasileiro de futebol da era de pontos corridos, desde de seu início (2003), até o último realizado (2023). Os *datasets* estão disponíveis no repositório do projeto, em [link]

A estrutura completa dos *datasets* é composta dos seguintes campos, auto-explicativos: *ano\_campeonato, data, rodada, estadio, arbitro, publico, publico\_max, time\_mandante, time\_visitante, tecnico\_mandante, tecnico\_visitante, colocacao\_mandante, colocacao\_visitante, valor\_equipe\_titular\_mandante, valor\_equipe\_titular\_visitante, idade\_media\_titular\_mandante, idade\_media\_titular\_visitante, gols\_mandante, gols\_visitante, gols\_1\_tempo\_mandante, gols\_1\_tempo\_visitante, escanteios\_mandante, escanteios\_visitante, faltas\_mandante, faltas\_visitante, chutes\_bola\_parada\_mandante, chutes\_bola\_parada\_visitante, defesas\_mandante, defesas\_visitante, impedimentos\_mandante, impedimentos\_visitante, chutes\_mandante, chutes\_visitante, chutes\_fora\_mandante, chutes\_fora\_visitante.*

# 3. Preparação da Base de Dados

Etapa de maior investimento de tempo e esforço..... [referências]

Xxxxxxxx

Em ML, o tempo gasto é muito maior no prepardo dos dados (análise exploratória) do que na modelagem em si,

# pois os modelos/algorítmos estão todos prontos nas bibliotecas, tendo somente que ter a poposta definida

# para aplicação. Mas não vamos gastar tempo escrevendo algorítmos.

# Mas temos também que selecionar bem os atributos que serão parâmetros de entrada

Pode-se notar que nem todos os atributos do *dataset* (listados no tópico 2.3) serão de nosso interesse, e ainda, que poderemos ter a necessidade criar algum campo de interesse, que venha a enriquecer a nossa base, e consequentemente, as nossas análises.

Nessa etapa, iniciamos com um entendimento geral da base de dados, para em seguida realizar a limpeza e transformações necessárias nos dados.

# 3.1. Ambiente Tecnológico (????)

Emm nosso projeto utilizamos...., em todas as etapas

A partir daqui utilizaremos python, google colab, .......

## 3.1. Análise Exploratória dos Dados

O objetivo da Análise Exploratória é estudar a base de dados que se está trabalhando e entender padrões e pontos de atenção que podem impactar em análises posteriores. [refer. curso python para data science]

Aqui nessa etapa, exploramos livremente o nosso *dataset*, para obter um maior entendimento do mesmo, e assim para “ganhar intimidade com os dados”. Aqui já começamos a entender as estratégias a serem adotadas para a efetiva preparação da base de dados, necessária antes de se aplicar qualquer algoritmo de *machine learning* sobre a base de dados.

Dentre as principais atividades realizadas na análise exploratória estão:

* Verificação da estrutura do *dataset* (atributos disponíveis);
* Verificação do tipo de dados dos atributos;
* Exploração visual dos dados;
* Obtenção da estatística descritiva dos atributos numéricos;
* Obtenção das tabelas de frequência dos atributos categóricos;
* Detecção de dados nulos;
* Detecção de outliers.

Realizamos nossa análise exploratória, através do notebook Python acessível em [link notebook], no qual estudamos de forma geral a estrutura e conteúdo da nossa base de dados.

Iniciamos realizando a leitura do *dataset*, gerando um *Dataframe* para a visualização de todas as colunas disponíveis. Em seguida, já selecionamos somente as colunas que, a princípio, seriam de nosso interesse. Na figura XX, é exibido um trecho do código onde selecionamos as colunas desejadas, e exibimos algumas linhas do *Dataframe* para análise inicial. Cada linha do *Dataframe* se refere a um jogo realizado.

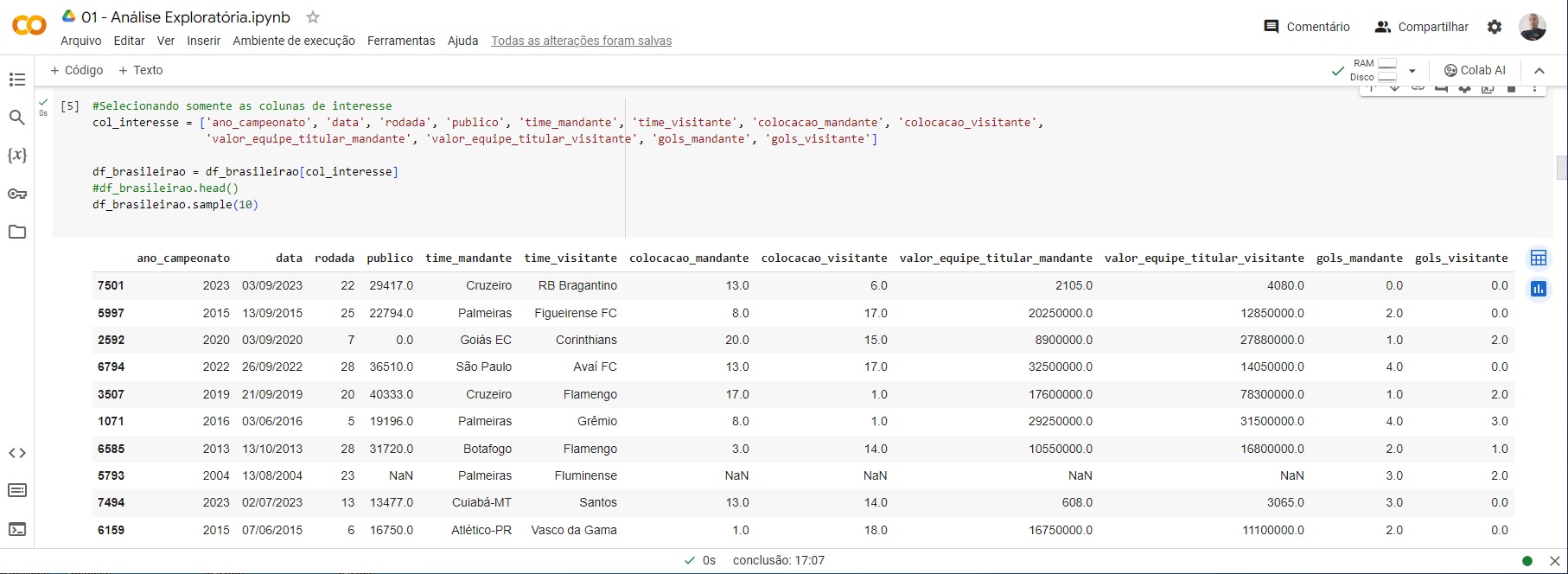


FIG. XX – Seleção das colunas de interesse e exibição do *Dataframe*

Após o reconhecimento da estrutura da base de dados, utilizamos a biblioteca **sweetviz**, a qual nos fornece uma análise exploratória geral, de forma automatizada. Parte dos resultados é exibida na figura XX.

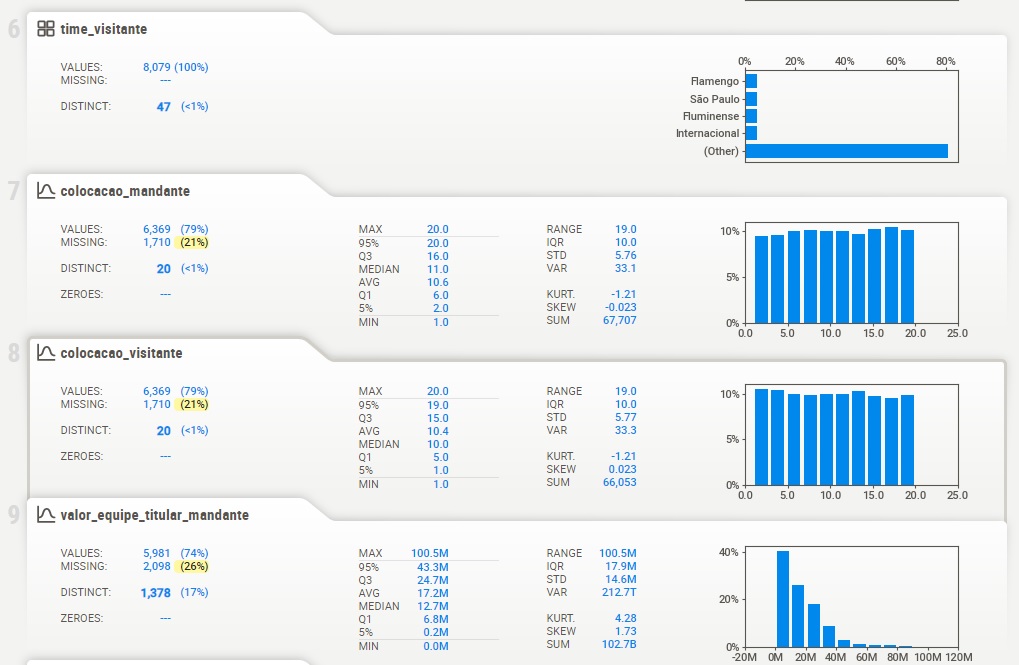


FIG. XX – Análise Exploratória geral com a biblioteca Sweetviz

Com a análise exploratória acima, já podemos observar que alguns atributos apresentam dados faltantes (nulos). Neste momento, estamos apenas realizando um conhecimento dos dados disponíveis. No entanto, mais à frente no projeto, quando efetivamente realizarmos a preparação da base de dados, vamos analisar individualmente cada atributo que apresenta dados vazios (nulos), para decidir a estratégia de preenchimento dos mesmos, ou ainda pela exclusão das linhas que apresentam atributos com dados nulos.

Outra percepção que obtivemos nesta etapa de análise exploratória, foi com relação aos atributos categóricos ***time\_mandante*** e ***time\_visitante***, onde foi possível observar em suas tabelas de frequência, a ocorrência duplicada dos seguintes times, os quais aparecem com nomes diferentes em linhas diversas do *Dataframe*:

* Athletico-PR e Atlético-PR
* Goiás e Goiás EC
* Santos e Santos FC

Assim, ao realizar a preparação da base de dados, teremos que padronizar esses nomes.

Para finalizar esta etapa de análise exploratória dos dados, observamos também o *dataset* **brasileirao\_serie\_a\_dados\_complementares\_2023.csv**, o qual foi preenchido por nós, com dados complementares (jogos no período de 27/10/2023 a 7/12/2023), os quais não estão disponibilizados no *dataset* **brasileirao\_serie\_a.csv**, conforme explicado no item 2.3 (Escolha e Obtenção das Bases de Dados) deste trabalho. Dessa forma, já sabíamos que os campos ***valor\_equipe\_titular\_mandante*** e ***valor\_equipe\_titular\_visitante***, estavam nulos, pois não tínhamos essa informação disponível para preenchimento. Ao realizarmos a preparação da base de dados, na próxima etapa do projeto, iremos decidir por uma estratégia de preenchimento dos mesmos.

A figura XX, a seguir ilustra um trecho do código, onde é realizada a leitura do *dataset* complementar, e exibidas algumas linhas do mesmo em um *dataframe*, onde pode-se observar os dados vazios dos atributos citados.

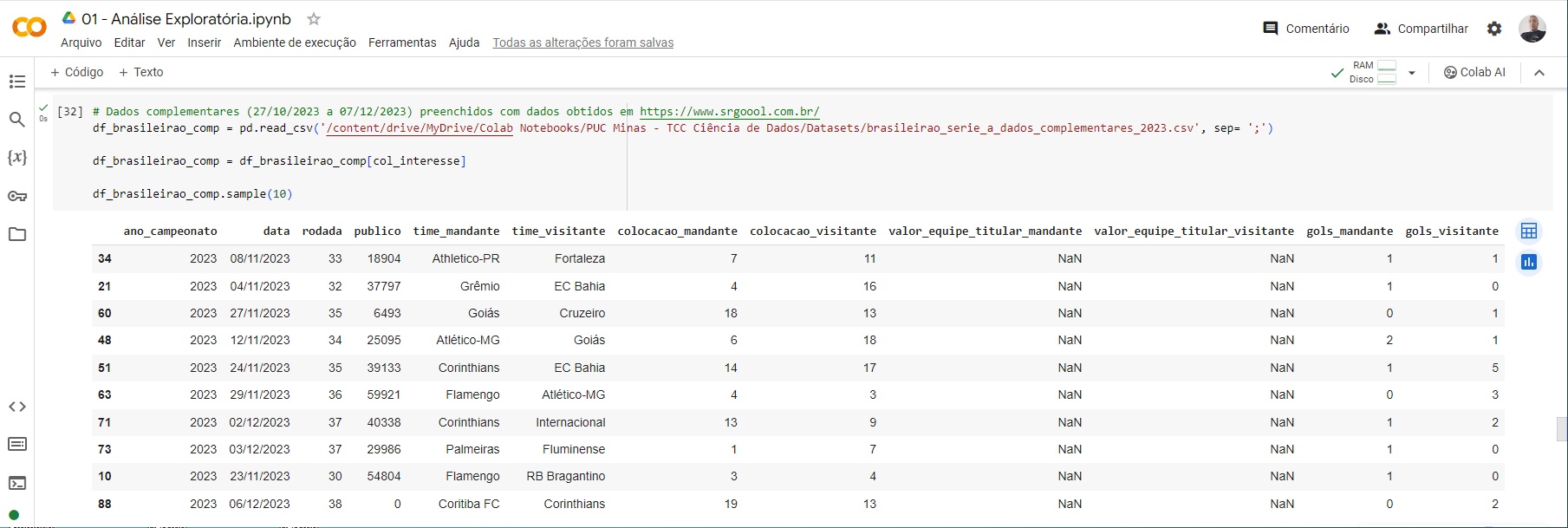


FIG. XX – *Dataframe* com os dados complementares

O Notebook Python completo com a Análise Exploratória tratada neste tópico pode ser acessado no repositório do projeto, em [link notebook].

Nos próximos tópicos, iremos avançar na preparação da base de dados, a partir das observações já realizadas até aqui, e realizaremos efetivamente os ajustes necessários na base de dados, deixando-a pronta para aplicação dos algoritmos de *machine learning*. Para fins de melhor organização, iniciaremos um novo Notebook Python onde concentraremos a preparação dos dados, senod o mesmo acessível em [link].

## 3.2. Seleção de Atributos – Redução de Dimensionalidade

## 3.3. Ajustes de Valores

## 3.3. Análise de Valores Vazios (Nulos)

## 3.4. Análise de Outliers

## 3.5. Engenharia de Atributos (feature enginering)

## (Criação de Novos Atributos)

## 3.6 Transformação de variáveis numéricas em categóricas

## 3.7 Agrupamento de categorias (variáveis categóricas)

## 3.8 Nova Seleção de Atributos

## 3.9. Padronização/Normalização de Variáveis

**(ajuste de escalas)**

## 3.10. Codificação de Variáveis Categóricas

# 4. Escolha do Modelo

(BOTAR GRÁFICO DA RETA DE REGRESSÃO DO MODELO ESCOLHIDO, PARA DADOS DE TREINAMENTO E DE TESTES – UM PRA CADA)

# 5. Implantação da Solução em Ambiente de Produção

# 6. Links

Os artefatos produzidos no desenvolvimento deste trabalho, bem como os *scripts*, *Datasets* e demais arquivos relevantes, estão disponíveis para acesso no repositório do projeto, no link [Relatório Técnico – Módulo A](https://github.com/salomaofreitasjr/TCC-PROJETO-INTEGRADO---PUC-MINAS---ANALYTICS-E-BI/tree/main/Projeto%20Integrado%20-%20Produ%C3%A7%C3%A3o/M%C3%B3dulo%20A).

# REFERÊNCIAS

[1.Sharda, Ramesh] SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business Intelligence e Análise de Dados para a Gestão do Negócio, 4ª Edição.** Porto Alegre: Bookman, 2019.

[2.Sarkar, Dipanjan] DIPANJAN, Sarkar; BALI, Raghav; SHARMA, Tushar. **Practical Machine Learning with Python.** Bangalore, Karnataka, India: Apress, 2018.

[3.Guido, Sarah] GUIDO, Sarah; MÜLLER, Andreas. **Introduction to Machine Learning with Python.** Sebastopol, CA, USA: O’Reilly, 2017.

[4.Faceli, Katti] FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André Carlos. **Inteligência Artificial, Uma Abordagem de Aprendizado de Máquina.** Rio de Janeiro : LTC, 2011.

[5.Gomes, Rodrigo] GOMES, Rodrigo. **Notas de aula da Disciplina Introdução à Linguagem Python, do Curso de Pós-Graduação em Analytics e Business Intelligence da PUC Minas.** Belo Horizonte, 2021.

[6.Gomes, Rodrigo] GOMES, Rodrigo. **Notas de aula da Disciplina Programação para Ciência de Dados, do Curso de Pós-Graduação em Analytics e Business Intelligence da PUC Minas.** Belo Horizonte, 2021.

[7.Fujii, Juliene] FUJII, Juliene Borges. **Notas de aula da Disciplina Estatística Geral – Teoria e Aplicações, do Curso de Pós-Graduação em Analytics e Business Intelligence da PUC Minas.** Belo Horizonte, 2021.

[8.De Paula, Hugo] DE PAULA, Hugo. **Notas de aula da Disciplina Machine Learning, do Curso de Pós-Graduação em Analytics e Business Intelligence da PUC Minas.** Belo Horizonte, 2021.

[9.Zárate, Luis] ZÁRATE, Luis E. **Notas de aula da Disciplina Modelagem e Preparação de Dados para Aprendizado de Máquina, do Curso de Pós-Graduação em Ciência de Dados e Big Data da PUC Minas.** Belo Horizonte, 2023.

[10.Patrocínio, Zenilton] PATROCÍNIO, Zenilton K. G. **Notas de aula da Disciplina Redes Neurais e Aprendizagem Profunda, do Curso de Pós-Graduação em Ciência de Dados e Big Data da PUC Minas.** Belo Horizonte, 2023.

[11.Alves, Pedro] ALVES, Pedro. **Apostila do Curso Python para Data Science e Analytics – Módulo Avançado, Aula 3 – Árvore de Decisão [https://python-para-datascience-analytics.club.hotmart.com/].** PA Analytics [https://paanalytics.net/], 2022.

[curso flai]

[curso pyhton impressionador]

(SUBSTITUIR PELAS FONTES DOS DATASETS DESTE PROJETO)

[16.Duque, Adão]. DUQUE, Adão. **Fontes de Dados do Campeonato Brasileiro de Futebol.** <https://www.kaggle.com/datasets/adaoduque/campeonato-brasileiro-de-futebol/metadata>. <https://github.com/adaoduque/Brasileirao_Dataset>. Acessado em 04/07/2022.