Universidade de Brasília

Faculdade UnB Gama

FGA0083 - Aprendizado de Máquina	**Semestre**: 2025.1
Docente: Sérgio Antônio Andrade	**Turma:** 01
Grupo 05	
Integrantes: Harryson Campos Martins Pedro Henrique Muniz de Oliveira Flávio Gustavo Araújo de Melo Leandro de Almeida Oliveira Jefferson Sena Oliveira José André Rabelo Rocha	**Matrícula:** 211039466 200059947 211030602 211030827 200020323 211062016

Mini Trabalho 4:

Preparação dos dados

Introdução

Este trabalho tem como foco a preparação dos dados para aplicação de modelos de aprendizado de máquina. Serão realizadas etapas como tratamento de valores ausentes, correção de outliers, codificação de variáveis categóricas e normalização dos dados. Essas etapas são essenciais para garantir a qualidade do dataset e a precisão dos modelos a serem desenvolvidos.

Objetivo

Preparar os dados coletados para modelagem de aprendizado de máquina, realizando limpeza, tratamento de valores ausentes, correção de outliers, codificação de variáveis e normalização, garantindo a qualidade do dataset e a eficácia dos modelos.

Código Utilizado

Todas as etapas descritas a seguir foram implementadas em um único código Python denominado limpeza_e_padronizacao_dos_dados.py, responsável por todo o processo de preparação dos dados nos três conjuntos analisados.

Tratamento de Valores Ausentes

Todos os valores ausentes encontrados nos datasets foram substituídos pela palavra "ausente". Essa padronização visa manter a consistência dos dados e evitar erros nas etapas seguintes de análise.

Padronização dos Nomes dos Times

Foi criado um dicionário com variações identificadas nos nomes dos times, que foi utilizado no código para uniformizar a nomenclatura das equipes em todos os conjuntos de dados.

Padronização da Data dos Jogos

As datas dos jogos foram convertidas para o formato YYYY-MM-DD em todos os conjuntos de dados analisados. Essa padronização facilita a ordenação cronológica e a integração das informações entre os conjuntos.

Remoção de outliers

Para a análise de dados do Campeonato Brasileiro de Futebol, a remoção dos outliers seria inadequada uma vez que estes valores extremos, como o América-RN com apenas 10,5% de aproveitamento ou times como Palmeiras e Grêmio com aproveitamento superior a 57% como mandantes, representam fenômenos reais e relevantes do futebol brasileiro (dominância histórica de certos clubes e as dificuldades enfrentadas por times recém-promovidos). Diferentemente de contextos onde outliers podem indicar erros de medição ou eventos anômalos a serem descartados, no futebol estes valores extremos trazem informações valiosas sobre a dinâmica do campeonato, a distribuição desigual de recursos entre clubes e os ciclos de sucesso e fracasso que caracterizam o esporte. Além disso, são essenciais para modelos preditivos realistas, e sua exclusão tornaria a análise artificialmente homogênea, distorcendo a dinâmica do campeonato.

Codificação de Variáveis Categóricas

As variáveis categóricas com até 20 categorias únicas foram transformadas em variáveis numéricas por meio de One-Hot Encoding. Essa transformação é necessária para que algoritmos de machine learning consigam processar variáveis de texto, convertendo-as em representações binárias que preservam a informação sem introduzir ordens artificiais entre as categorias.

Padronização de Variáveis Numéricas

As variáveis numéricas selecionadas foram padronizadas utilizando o método de Z-score, que ajusta os valores para uma distribuição com média igual a zero e desvio padrão igual a um. Essa padronização é fundamental para garantir que variáveis com diferentes escalas tenham o mesmo peso em algoritmos sensíveis à magnitude dos dados, como regressão logística, SVM e KNN, favorecendo uma modelagem mais equilibrada e precisa.

Conclusão

A preparação dos dados seguiu um fluxo focado na integridade, consistência e representatividade do fenômeno estudado. As etapas de tratamento de valores ausentes, padronização de nomenclaturas, normalização de datas, codificação de variáveis categóricas e padronização de variáveis numéricas foram aplicadas de forma a preservar as características essenciais do Campeonato Brasileiro de Futebol, respeitando tanto as particularidades dos dados quanto às exigências dos modelos analíticos. Com essas transformações, o dataset encontra-se devidamente estruturado, coeso e pronto para a aplicação de técnicas de modelagem preditiva,

assegurando análises mais robustas, interpretáveis e alinhadas à realidade do contexto esportivo estudado.