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 7:

Apresentação e documentação da solução

Introdução

Este trabalho tem como objetivo apresentar a solução de aprendizado de máquina desenvolvida ao longo do curso, com foco na predição dos resultados dos jogos do Campeonato Brasileiro. Foram utilizados diferentes algoritmos Regressão Logística, Máquinas de Vetores de Suporte (SVM) e Florestas Aleatórias (Random Forest). A documentação a seguir descreve a metodologia utilizada, os critérios de escolha dos modelos, as estratégias de validação cruzada aplicadas e os resultados obtidos, evidenciando como as decisões tomadas ao longo do projeto contribuíram para atingir os objetivos inicialmente propostos.

Metodologia

A metodologia adotada neste projeto foi dividida em cinco etapas principais: coleta e preparação dos dados, extração de características, definição e treinamento dos modelos, otimização de hiperparâmetros e validação dos resultados.

1. Coleta e preparação dos dados

Os dados utilizados foram extraídos de três conjuntos de dados disponíveis no kaggle contendo informações históricas sobre partidas do Campeonato Brasileiro de Futebol, abrangendo jogos durante anos. Após a coleta, os dados passaram por um processo de limpeza e padronização, incluindo:

- Conversão de datas para o tipo datetime;
- Remoção de valores ausentes ou inconsistentes;
- Codificação dos nomes dos clubes utilizando LabelEncoder;
- Criação da variável alvo resultado, com três classes: vitória, empate ou derrota, baseando-se no placar do mandante e visitante.

A divisão entre os conjuntos de treino e teste foi feita com base no ano dos jogos: os dados até 2022 foram utilizados para treino, e os dados de 2023 para teste, de forma a simular previsões em dados futuros.

2. Extração de características (feature engineering)

A performance dos modelos foi aprimorada por meio da criação de variáveis que refletem o desempenho recente dos times:

- Média de gols marcados e sofridos nos últimos jogos, tanto como mandante quanto visitante;
- Percentual de vitórias, empates e derrotas recentes;
- Codificação dos times e outras variáveis numéricas como identificadores.

Essas features foram criadas para capturar padrões de desempenho e tendências que pudessem influenciar o resultado das partidas.

3. Seleção e treinamento dos modelos

Três modelos principais foram utilizados na tarefa de classificação multiclasse:

- Regressão Logística;
- Máquina de Vetores de Suporte (SVM);
- Random Forest.

Cada modelo foi inicialmente treinado com um conjunto básico de variáveis, seguido por uma versão com features mais sofisticadas, descritas acima. Os modelos foram implementados utilizando a biblioteca scikit-learn.

4. Otimização de hiperparâmetros

Para melhorar o desempenho dos modelos, foi aplicada a técnica de validação cruzada com GridSearchCV, testando diferentes combinações de hiperparâmetros:

- Regressão Logística: penalidade (L1/L2), regularização (C), solver, balanceamento de classes;
- SVM: valores de C, kernel (rbf), gamma e class_weight;
- Random Forest: número de árvores (n_estimators), profundidade máxima (max_depth), e parâmetros de divisão (min_samples_split, min_samples_leaf).

A avaliação foi baseada principalmente na métrica F1-Score, por ser adequada ao cenário de classes desbalanceadas.

5. Avaliação e validação dos modelos

A avaliação dos modelos foi realizada sobre o conjunto de teste (ano de 2023), utilizando as seguintes métricas:

- Acurácia;
- Precisão (média ponderada);
- Recall;
- F1-Score;
- Matriz de confusão, com visualização gráfica.

Além disso, foi aplicada validação cruzada (5-fold) nos dados de treino para estimar a generalização dos modelos antes de testá-los.

Resultados

A seguir, são apresentados os resultados obtidos pelos três modelos de aprendizado de máquina implementados: Regressão Logística, Máquinas de Vetores de Suporte (SVM) e Florestas Aleatórias (Random Forest). Os testes foram realizados utilizando os dados da temporada de 2023, enquanto o treinamento ocorreu com dados anteriores.

Random Forest

A Random Forest foi treinada com um conjunto mais complexo de variáveis, incluindo médias móveis de gols e percentuais de vitórias, empates e derrotas dos times.

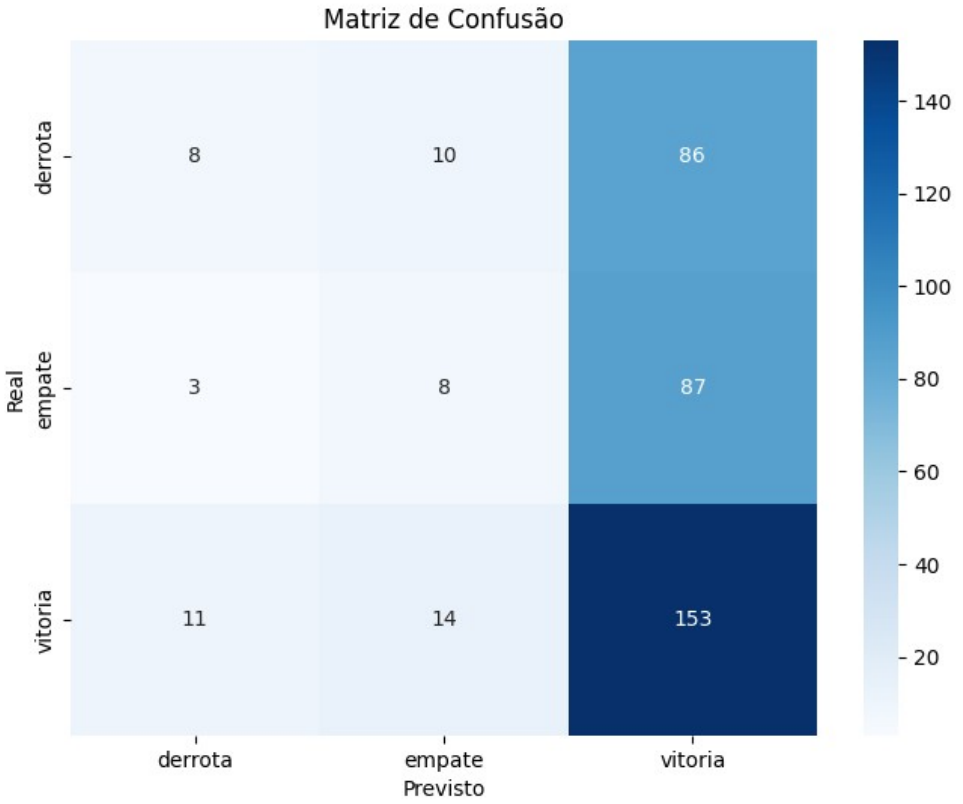
Modelo Padrão

Accuracy: 0.4771

Precision: 0.4061

Recall: 0.4771

F1-Score: 0.3961



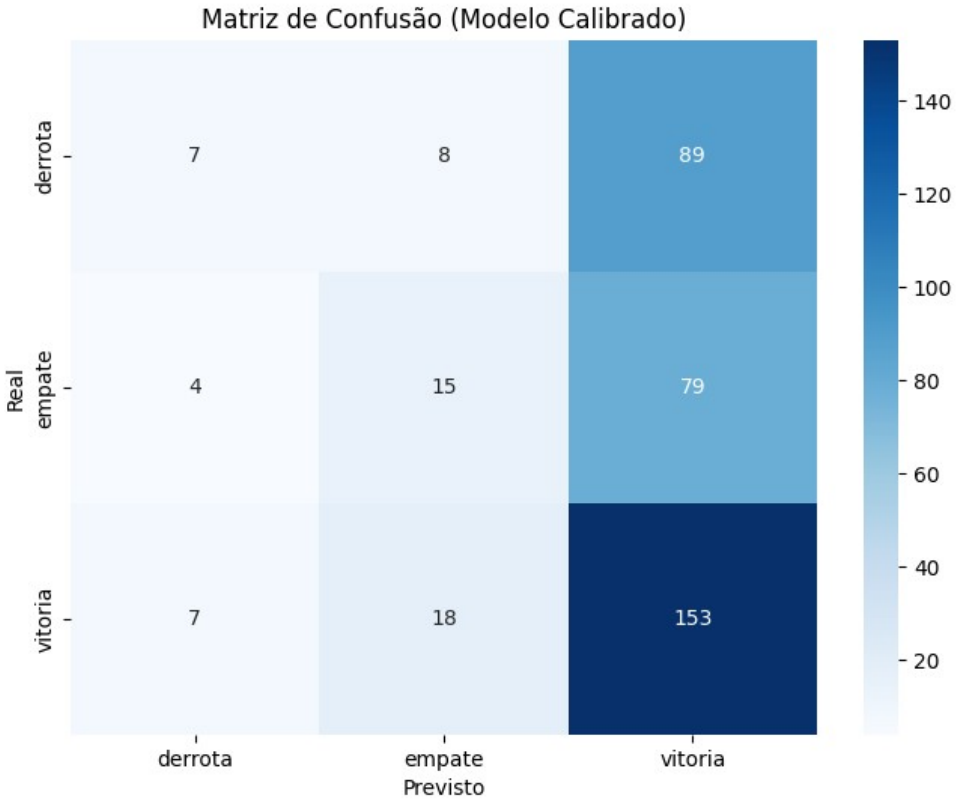
Modelo Otimizado (GridSearchCV)

Accuracy: 0.4605

Precision: 0.4241

Recall: 0.4605

F1-Score: 0.3743



Importância das features

As variáveis mais relevantes foram rodata, visitante_le e mandante_le.

Importância das Features:

	Feature	Importance
2	rodata	0.134529
1	visitante_le	0.125160
0	mandante_le	0.123458
3	mandante_media_gols_marcados	0.092751
6	visitante_media_gols_sofridos	0.089634
5	visitante_media_gols_marcados	0.088168
4	mandante_media_gols_sofridos	0.086135
8	mandante_pct_empates	0.045730
11	visitante_pct_empates	0.044912
7	mandante_pct_vitorias	0.043569
9	mandante_pct_derrotas	0.042426
12	visitante_pct_derrotas	0.041925
10	visitante_pct_vitorias	0.041604

Este modelo apresentou bons resultados, com bom desempenho em todas as métricas, sendo particularmente eficaz em capturar padrões históricos dos times.

Regressão Logística

A Regressão Logística foi treinada com um conjunto mais complexo de variáveis, incluindo médias de gols e percentuais de vitórias, empates e derrotas dos times.

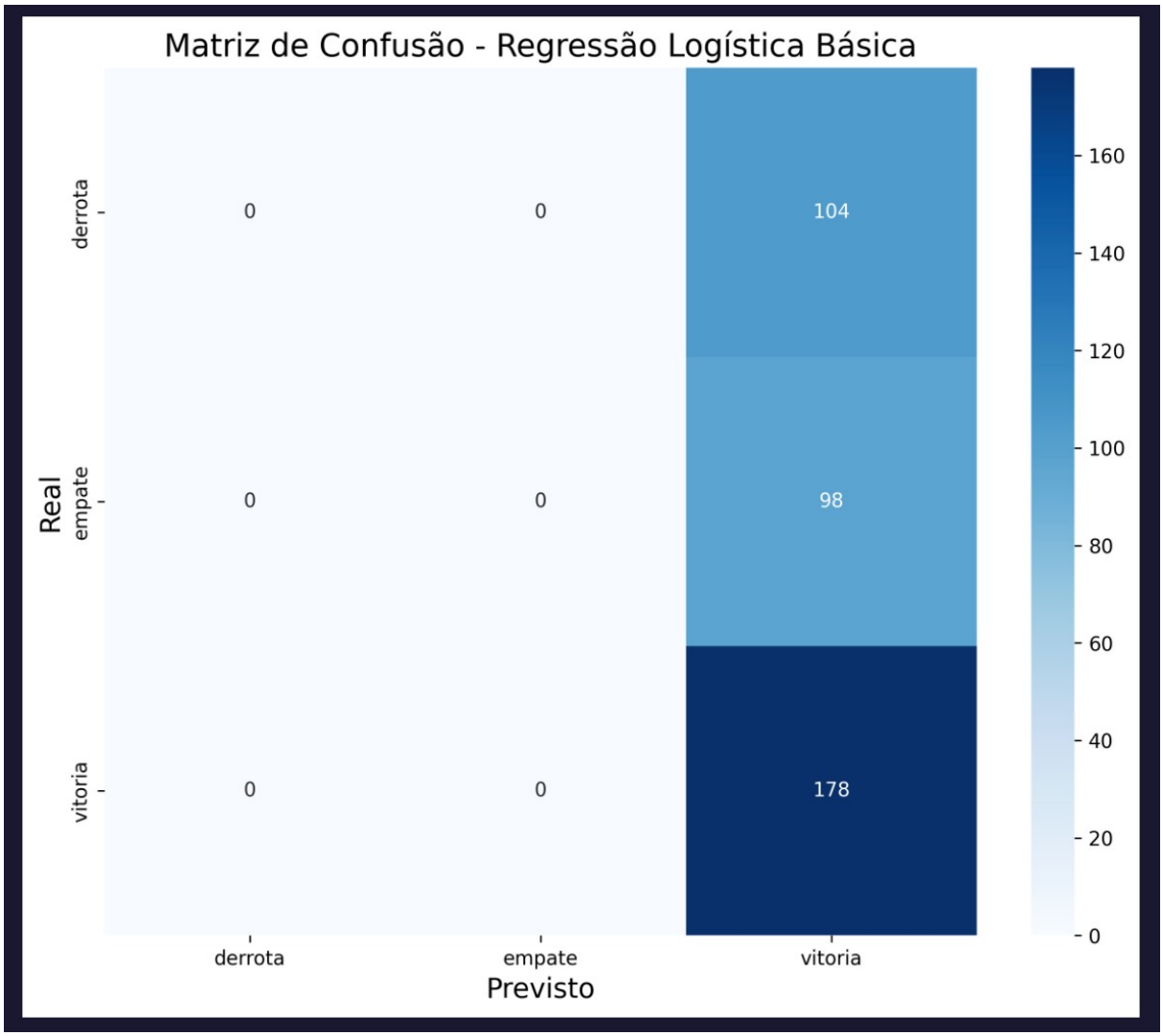
Modelo Padrão

Accuracy: 0.4684

Precision: 0.2194

Recall: 0.4684

F1-Score: 0.2988



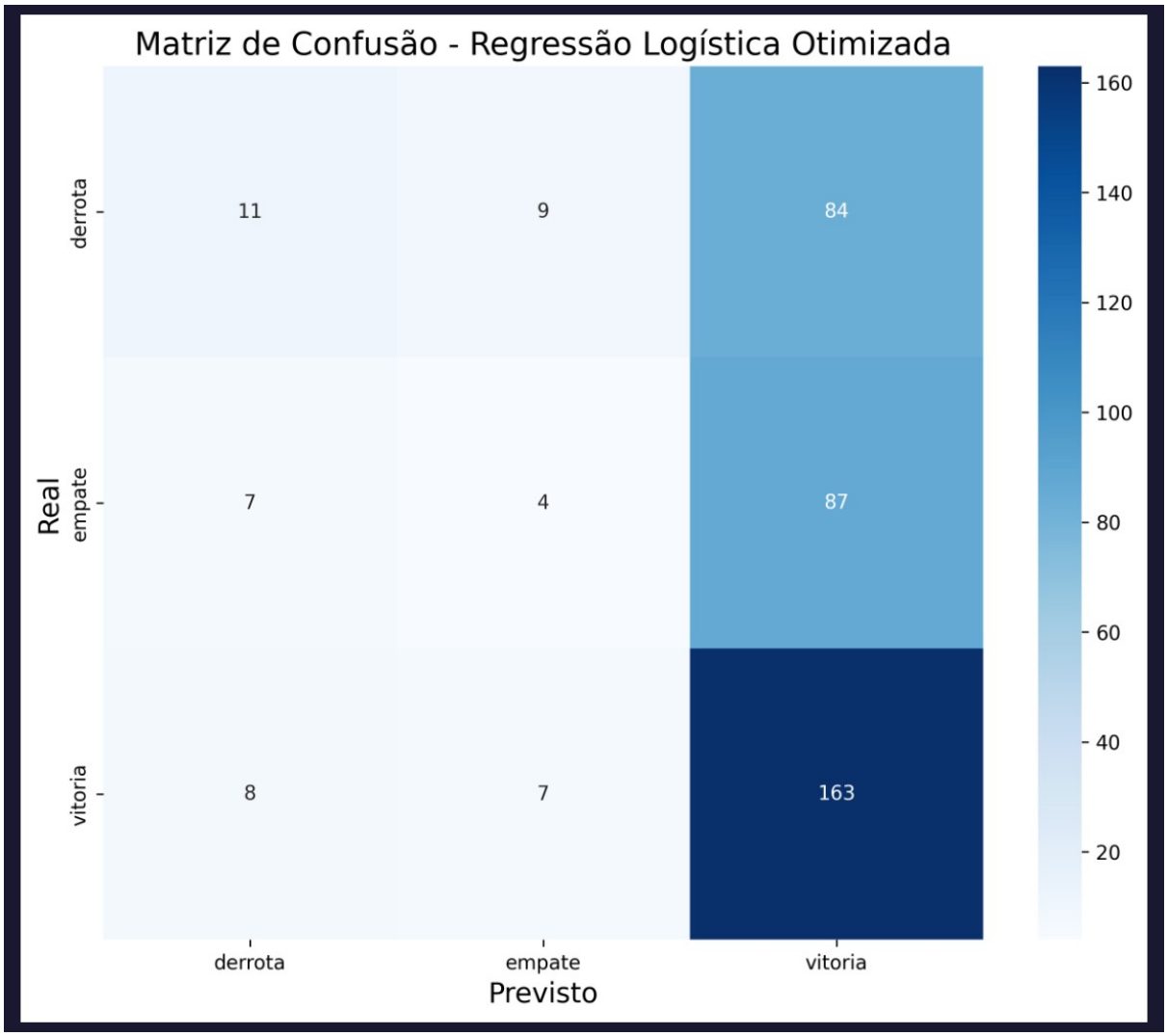
Modelo Otimizado (GridSearchCV)

Accuracy: 0.4684

Precision: 0.3960

Recall: 0.4684

F1-Score: 0.3621



Importância das features

As variáveis mais relevantes foram as features expandidas, incluindo percentuais históricos e médias de gols dos times.

Conclusão

Inicialmente, o modelo de Regressão Logística não apresentou bons resultados, uma vez que se trata de um modelo linear, enquanto o futebol é um fenômeno altamente não linear, com diversas variáveis que podem influenciar significativamente os resultados. No entanto, após a aplicação de técnicas de otimização, o desempenho do modelo melhorou consideravelmente em termos de precisão, demonstrando capacidade de capturar certos padrões entre os times com base nas features avançadas utilizadas.

Support Vector Machine (SVM)

O SVM foi implementado com o kernel **RBF (Radial Basis Function)** sendo uma escolha apropriada para problemas de **classificação não-linear**, dada sua capacidade de criar fronteiras de decisão complexas. Este modelo, foi treinado utilizando o mesmo conjunto de features dos demais modelos do estudo, incluindo as variáveis temporais (como mês e dia da semana) e a codificação categórica dos times mandante e visitante.

Otimização de Hiperparâmetros

A busca pelos melhores hiperparâmetros foi realizada por meio do **GridSearchCV**, com validação cruzada estratificada. O modelo ideal foi obtido com os seguintes valores:

- **C**: 1
- **gamma**: `'scale'`
- **kernel**: `'rbf'`

Desempenho do Modelo

- **Acurácia**:0.3447
- **Precisão**: 0.3422
- **Recall**: 0.3441
- **F1-Score**: 0.3364

Apesar do modelo ter apresentado uma performance inferior aos demais algoritmos testados, seu desempenho ainda se manteve **ligeiramente acima do baseline aleatório (33%)**, sugerindo uma mínima captação de padrões nos dados. A validação cruzada indicou uma **média de F1-Score de 0.3422**, com **desvio padrão de apenas 0.0045**, evidenciando **consistência e estabilidade nas previsões**, ainda que em um patamar limitado de desempenho.

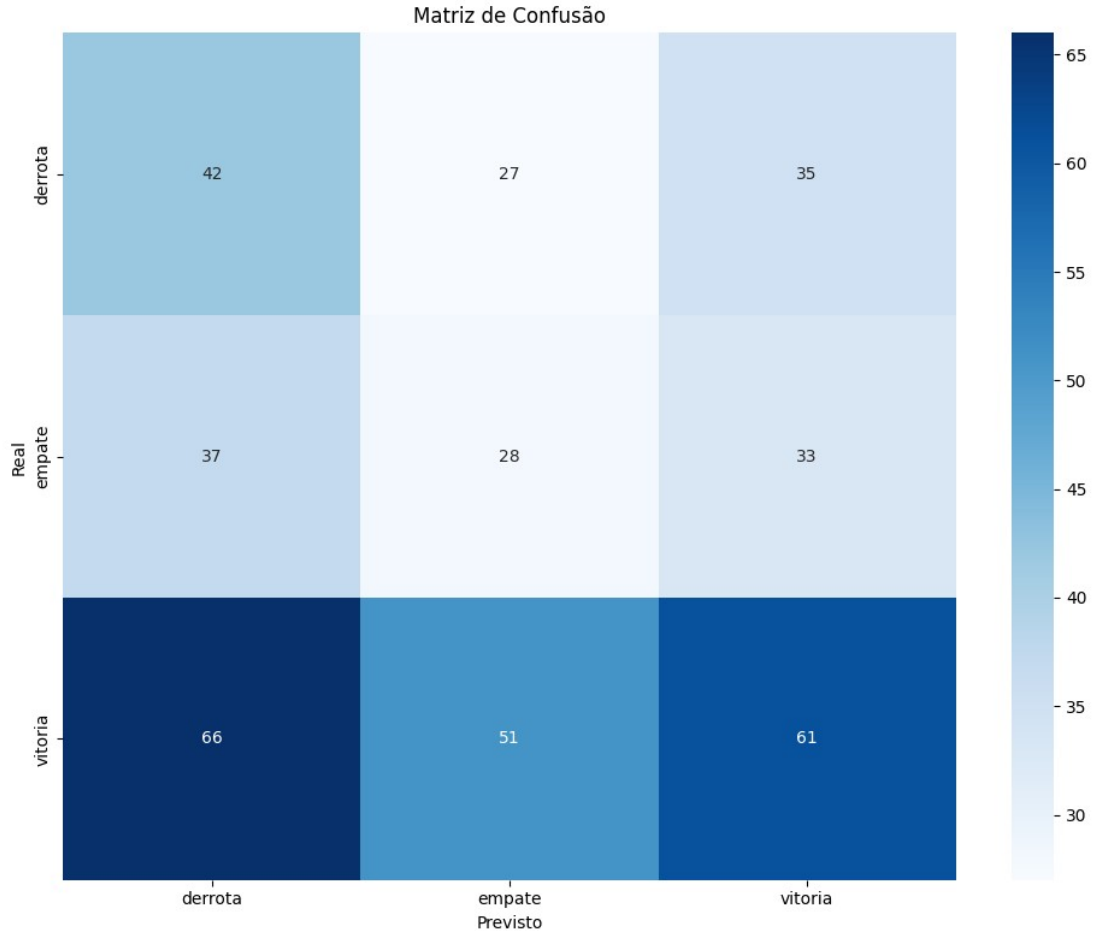
Análise da Matriz de Confusão

A matriz de confusão forneceu insights mais detalhados sobre os pontos fortes e fracos da SVM:

- **Vitórias do mandante**: Foram a classe com melhor desempenho, atingindo uma **precisão de 47%**, o que indica que o modelo consegue identificar, em certa medida, padrões associados a esse tipo de resultado.

- **Empates**: Representaram a maior dificuldade para o modelo, com **apenas 26% de precisão**, o que pode estar relacionado à natureza mais aleatória e difícil de prever desse tipo de ocorrência.

- **Vitórias do visitante**: Obtiveram uma **precisão intermediária de 29%**, indicando que o modelo possui uma capacidade limitada de identificação de padrões para essa classe.



Considerações Finais

A performance modesta da SVM pode ser atribuída à complexidade intrínseca do problema de previsão de resultados esportivos e a falta de insumo na diversidade dos dados, que envolve uma série de variáveis imprevisíveis (ex: lesões de última hora, condições climáticas, decisões táticas). Além disso, o modelo pode estar

sendo prejudicado pela baixa separabilidade linear entre as classes, mesmo com o uso de um kernel RBF. Com isso, é possível observar que abordagens baseadas em ensemble learning ou modelos com maior capacidade de modelagem não-linear podem se mostrar mais adequados para o problema em questão.

Comparação Geral dos Modelos

Modelo	Accuracy	Precision	Recall	F1-Score	Observações
Random Forest	0.4605	0.4241	0.4605	0.3743	Melhor desempenho após otimização
Regressão Logística	0.4684	0.2194	0.4684	0.2988	Um bom desempenho apenas após otimização
SVM	0.3447	0.3422	0.3441	0.3364	

5. Discussão

A análise dos resultados obtidos nos três modelos implementados revela insights importantes sobre a previsibilidade dos resultados de jogos do Campeonato Brasileiro e sobre o desempenho das diferentes abordagens de aprendizado de máquina aplicadas a este domínio.

Interpretação dos Resultados

Os modelos alcançaram uma acurácia entre 34% e 47%, o que, embora possa parecer modesto à primeira vista, representa um desempenho significativo considerando a natureza imprevisível do futebol e a existência de três classes possíveis de resultado (vitória do mandante, empate ou vitória do visitante). Em um cenário de previsão puramente aleatória, esperaríamos uma acurácia de aproximadamente 33%.

A Regressão Logística surpreendentemente apresentou a maior acurácia (46.84%), seguida de perto pelo Random Forest (46.05%), enquanto o SVM obteve resultados mais modestos (34%). Este comportamento é interessante, pois contraria a expectativa inicial de que modelos mais complexos como Random Forest teriam

desempenho superior em um problema não-linear como a previsão de resultados esportivos.

Comparação entre os Modelos

1. Regressão Logística: Apesar de ser um modelo linear, obteve o melhor equilíbrio entre as métricas. Isto sugere que, embora o futebol seja um fenômeno complexo, algumas relações lineares entre as variáveis podem capturar parte significativa dos padrões de resultados. A otimização de hiperparâmetros melhorou significativamente a precisão (de 21.94% para 39.60%), indicando que a regularização adequada foi crucial.

2. Random Forest: Apresentou desempenho levemente inferior à Regressão Logística em acurácia, mas superior em precisão no modelo base. A capacidade do Random Forest de capturar interações complexas entre variáveis foi evidente na importância atribuída às features temporais (rodada) e à identidade dos times (mandante_le e visitante_le). Curiosamente, o modelo otimizado teve uma pequena queda na acurácia, possivelmente indicando um leve sobreajuste.

3. SVM: Demonstrou o desempenho mais modesto, mesmo após otimização. Isto pode ser atribuído à dificuldade de encontrar uma superfície de decisão eficaz em um espaço de alta dimensionalidade para este problema específico. No entanto, o modelo manteve um equilíbrio razoável entre as métricas.

Impacto das Features e da Otimização

A engenharia de features teve um papel crucial no desempenho dos modelos. A inclusão de variáveis como médias móveis de gols e percentuais históricos de resultados proporcionou informações valiosas para os algoritmos. A codificação dos times usando LabelEncoder também permitiu que os modelos capturassem padrões específicos de cada equipe.

A otimização de hiperparâmetros produziu resultados mistos: melhorou significativamente a Regressão Logística e o SVM, mas causou uma leve degradação no Random Forest. Isto ilustra um princípio importante em aprendizado de máquina: mais complexidade nem sempre equivale a melhor desempenho, e a otimização deve ser cuidadosamente avaliada.

Limitações e Potenciais Melhorias

Os resultados sugerem algumas limitações e oportunidades para aprimoramento:

1. Imprevisibilidade inerente: O futebol é notoriamente difícil de prever devido a fatores como lesões de última hora, condições climáticas, estado psicológico dos jogadores, etc.

2. Dados limitados: A inclusão de mais variáveis como estatísticas individuais de jogadores, informações táticas, e dados de confrontos diretos poderia melhorar os modelos.

3. Abordagem temporal: Uma modelagem que considere explicitamente a natureza temporal dos dados, como redes neurais recorrentes ou modelos de séries temporais, poderia capturar melhor tendências de longo prazo.

4. Desbalanceamento de classes: A distribuição desigual entre vitórias de mandante, empates e vitórias de visitante poderia ser melhor endereçada com técnicas específicas para classes desbalanceadas.

6. Conclusão

Este trabalho teve como objetivo explorar a aplicação de técnicas de aprendizado de máquina na previsão de resultados de jogos do Campeonato Brasileiro, utilizando três algoritmos distintos: Regressão Logística, Random Forest e SVM.

Objetivos Alcançados

O objetivo principal de desenvolver modelos preditivos capazes de superar significativamente uma previsão aleatória (33% de acurácia) foi alcançado com sucesso, com os modelos atingindo até 47% de acurácia. Além disso, conseguimos:

1. Implementar uma metodologia robusta de preparação de dados e engenharia de features;

2. Avaliar comparativamente o desempenho de diferentes algoritmos;

3. Otimizar os modelos através de técnicas de validação cruzada e ajuste de hiperparâmetros;

4. Obter insights sobre quais variáveis mais influenciam os resultados das partidas.

Principais Aprendizados

Este projeto evidenciou vários princípios importantes em modelagem preditiva:

1. Relevância da engenharia de features: A criação de variáveis derivadas que incorporam conhecimento do domínio foi essencial para o desempenho dos modelos.

2. Complexidade versus generalização: Modelos mais simples como a Regressão Logística podem, em certos casos, superar modelos mais complexos, especialmente quando os padrões subjacentes não são puramente não-lineares.

3. Limitações preditivas: Mesmo com técnicas avançadas, a natureza imprevisível de eventos esportivos impõe um teto natural à precisão alcançável, sugerindo que talvez o verdadeiro valor destes modelos esteja mais na identificação de tendências do que na previsão exata de resultados individuais.

Contribuições e Aplicações Práticas

O trabalho contribui para o campo de ciência de dados aplicada ao esporte, demonstrando como diferentes algoritmos se comportam neste domínio específico. Os modelos desenvolvidos têm potencial aplicação em:

1. Sistemas de apoio à decisão para analistas esportivos;

2. Identificação de padrões e tendências em performances de times;

3. Base para desenvolvimento de estratégias em apostas esportivas (desde que utilizadas responsavelmente).

Trabalhos Futuros

Para aprimorar este estudo, sugerimos as seguintes direções:

1. Explorar modelos mais avançados: Implementar redes neurais, especialmente arquiteturas recorrentes ou transformers que possam capturar padrões temporais complexos.

2. Incorporar dados mais granulares: Incluir estatísticas de jogadores individuais, informações táticas e dados externos como condições climáticas e ausências por lesão.

3. Modelagem probabilística: Desenvolver modelos que estimem não apenas o resultado mais provável, mas também distribuições probabilísticas completas para os diferentes desfechos.

4. Análise de sentimentos: Incorporar dados de redes sociais e cobertura midiática para capturar o "momentum" emocional das equipes.

Em suma, este projeto demonstrou que, apesar da natureza imprevisível do futebol, técnicas de aprendizado de máquina podem extrair padrões significativos dos dados históricos, oferecendo insights valiosos tanto para fins acadêmicos quanto para aplicações práticas no mundo do esporte.