UNIVERSIDADE ESTADUAL DO PIAUÍ
CENTRO DE TECNOLOGIA E URBANISMO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Ryan Rodrigo da Cruz Oliveira

Abordagem Baseada em Gradient Boosting para Previsão de Resultados em Corridas de Galgos

TERESINA
2025

Ryan Rodrigo da Cruz Oliveira

Abordagem Baseada em Gradient Boosting para Previsão de Resultados em Corridas de Galgos

Monografia de Trabalho de Conclusão de Curso apresentado na Universidade Estadual do Piauí – UESPI como parte dos requisitos para conclusão do Curso de Bacharelado em Ciência da Computação.

Orientador: Prof Dr. Sérgio Barros de Sousa

TERESINA

2025

O48a Oliveira, Ryan Rodrigo da Cruz.
Abordagem baseada em Gradient Boosting para previsão de resultados em corridas de Galgos / Ryan Rodrigo da Cruz Oliveira. - 2025.
53 f.: il.

Monografia (graduação) - Universidade Estadual do Piauí-UESPI, Bacharelado em Ciências da Computação, Campus Poeta Torquato Neto, Teresina-PI, 2025.
"Orientador: Prof. Dr. Sérgio Barros de Sousa".

1. Aprendizado de máquina. 2. Previsão de resultados. 3. Gradient Boosting. I. Sousa, Sérgio Barros de . II. Título.

CDD 004.07

Ficha elaborada pelo Serviço de Catalogação da Biblioteca da UESPI
Nayla Kedma de Carvalho Santos (Bibliotecário) CRB-3ª/1188

Abordagem Baseada em Gradient Boosting para Previsão de Resultados em Corridas de Galgos

Ryan Rodrigo da Cruz Oliveira

Monografia de Trabalho de Conclusão de Curso apresentado na Universidade Estadual do Piauí – UESPI como parte dos requisitos para conclusão do Curso de Bacharelado em Ciência da Computação.

Prof Dr. Sérgio Barros de Sousa, Dsc.
Orientador

Nota da Banca Examinadora: 10,00

Banca Examinadora:

Documento assinado digitalmente
gov.br
SERGIO BARROS DE SOUSA
Data: 07/07/2025 12:05:51-0300
Verifique em https://validar.iti.gov.br

Prof Dr. Sérgio Barros de Sousa, Dsc.
Presidente

Documento assinado digitalmente
gov.br
CARLOS GIOVANNI NUNES DE CARVALHO
Data: 08/07/2025 11:57:13-0300
Verifique em https://validar.iti.gov.br

Dr. Carlos Giovanni Nunes de Carvalho, Dsc.
Membro

Documento assinado digitalmente
gov.br
Reginaldo Rodrigues das Gracas
Data: 07/07/2025 15:47:03-0300
Verifique em https://validar.iti.gov.br

Me. Reginaldo Rodrigues Das Graças, Dsc.
Membro

“A arma mais forte em dois milhões de anos
de história humana: tecnologia de comunicação.“
(Inagaki, Riichiro; Dr. Stone, 2019)

.

RESUMO

Corridas de galgos, um esporte tradicional no Reino Unido, atraem tanto entusiastas quanto apostadores que visam lucrar nesse ramo e em 2020 esse mercado movimentou mais de 230 milhões de libras esterlinas em apostas presenciais e fora dos locais de corrida. Enquanto muitos baseiam suas decisões apenas na intuição, apostadores mais experientes utilizam técnicas de análise do histórico dos galgos para tentar prever o vencedor, o que pode ser difícil para um iniciante devido à complexidade dos dados a serem analisados. Este trabalho investigou a aplicação de técnicas de aprendizado de máquina na previsão de resultados em corridas de galgos, desenvolvendo um modelo preditivo capaz de estimar o tempo de corrida dos galgos utilizando o algoritmo de regressão da biblioteca *CatBoost*, que permitiu projetar uma classificação e identificar estratégias de aposta nesse mercado. A análise envolveu a estruturação de uma *pipeline* de preparação de dados baseada em trabalhos relevantes na literatura atual e a utilização de técnicas de otimização de hiperparâmetros no treinamento do modelo. Para assegurar a capacidade de generalização para novos dados, foi realizada uma avaliação da precisão na previsão do tempo de corrida por meio de validação cruzada *k-fold*. O trabalho contribuiu para o progresso do conhecimento na criação de modelos preditivos em aprendizado de máquina aplicados a bases de dados heterogêneas e para a identificação de estratégias de apostas esportivas em corridas de galgos.

Palavras-chaves: Aprendizado de Máquina. Previsão de Resultados. Corridas de Galgos. *Gradient Boosting*.

ABSTRACT

Greyhound racing, a traditional sport in the United Kingdom, attracts both enthusiasts and bettors seeking profit in this field. In 2020, this market generated over £230 million in on-site and off-site betting. While many individuals rely solely on intuition, more experienced bettors employ techniques to analyze greyhounds' historical performance in an attempt to predict the winner, a task that can be challenging for beginners due to the complexity of the data involved. This study investigated the application of machine learning techniques for predicting outcomes in greyhound races, by developing a predictive model capable of estimating race times using the CatBoost regression algorithm. This allowed for the projection of race rankings and the identification of potential betting strategies within this market. The analysis involved the construction of a data preparation pipeline inspired by relevant works in the current literature, as well as the use of hyperparameter optimization techniques during model training. To ensure the model's ability to generalize to unseen data, prediction accuracy was evaluated through k-fold cross-validation. This study contributes to the advancement of knowledge in the development of predictive models in machine learning applied to heterogeneous datasets, and to the identification of sports betting strategies in greyhound racing.

Keywords: Machine Learning. Outcome Prediction. Greyhound Racing. Gradient Boosting.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ML	*Machine Learning* (Aprendizado de Máquina)
GBM	*Gradient Boosting Machine*
GIGO	*Garbage In, Garbage Out* (Entrada Lixo, Saída Lixo)
GBGB	*Greyhound Board of Great Britain* (Conselho de Galgos da Grã-Bretanha)
RMSE	*Root Mean Square Error* (Raiz do Erro Quadrático Médio)
MAE	*Mean Square Error* (Erro Médio Absoluto)
ORM	*Object-Relational Mapping*
LDA	*Linear Discriminant Analysis* (Análise Discriminante Linear)
XAI	*Explainable Artificial Intelligence* (Inteligência Artificial Explicável)
CART	*Classification and Regression Tree*
RF	*Random Forest*
NN	*Neural Network*
SVM	*Support Vector Machine*

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Representando um dos esportes mais tradicionais no Reino Unido, as corridas de galgos mantêm uma vasta base de entusiastas e apostadores. O esporte que já atraiu milhões de espectadores em estádios no século passado, enfrenta atualmente uma redução gradual de público presencial (Laybourn, 2019). Contudo, permanece economicamente relevante devido ao forte impacto do mercado de apostas. No Brasil, o mercado de apostas esportivas ganha força desde a regulamentação pela Lei n.º 13.756/2018, que estabeleceu diretrizes para a exploração comercial de apostas de quota fixa, ampliando as possibilidades de atuação no segmento (Gazeta do Povo, 2023).

No entanto, a previsão de resultados em corridas de galgos é desafiadora, principalmente devido à baixa qualidade e à heterogeneidade dos dados históricos. Problemas como valores ausentes, condições variáveis da pista, categorias de corrida e posições de largada complicam as análises preditivas (Lyons, 2016). Diferentemente de outros esportes, que possuem maior uniformidade nas condições de jogo, as corridas de galgos envolvem fatores voláteis e não padronizados. Nesse cenário, o aprendizado de máquina é uma ferramenta poderosa para análise de dados, capaz de identificar padrões em diversos tipos de dados, possibilitando a aplicação em diversos segmentos, inclusive o ramo esportivo (Sports, 2024). O aprendizado de máquina oferece uma ampla gama de algoritmos, cada um com suas particularidades e aplicações específicas.

Nesse contexto, a biblioteca *CatBoost* baseada em *gradient boosting*, é amplamente reconhecida por sua eficiência no tratamento de variáveis categóricas sem a necessidade de conversão para formatos numéricos como demonstrado por (Prokhorenkova et al., 2018), tornando o seu algoritmo de regressão *CatBoostRegressor* uma escolha promissora, o que permite estimar o tempo de corrida de cada galgo com base no histórico recente de competições. A previsão de tempo permitirá estabelecer uma classificação projetada, facilitando a aplicação de diferentes cenários de aposta presentes no mercado, além do vencedor da corrida, existem outras classificações, dentre as quais o *Top 3* que consiste no galgo finalizar entre as 3 primeiras posições e o *Lay Placed*, que caracteriza uma aposta contra na qual o galgo não pode alcançar a segunda posição.

Para que o *CatBoostRegressor* desempenhe todo o seu potencial na previsão dos tempos de corrida dos galgos, uma etapa essencial é o tratamento adequado dos dados a serem utilizados. Isso se deve ao fato de que a qualidade dos dados influencia

diretamente nos resultados obtidos, como destacado no estudo (Frye; Schmitt, 2020), assegurar a qualidade de dados necessária é considerado um dos maiores desafios, e a baixa qualidade de dados resulta em análises de baixa qualidade, o que também é conhecido como o princípio do *Garbage In, Garbage Out* (GIGO), conforme discutido por Wolff.

Neste cenário, o presente trabalho propõe explorar a aplicação de *machine learning* (ML) em corridas de galgos, selecionando o *CatBoostRegressor* para essa análise, com o objetivo de desenvolver um modelo para as previsões e analisar o desempenho nas apostas esportivas visando aperfeiçoar as análises e auxiliar apostadores do mercado.

1.1 Justificativa

A crescente aplicação de Machine Learning (ML) na previsão de resultados esportivos contrasta com uma notável lacuna na literatura científica, especificamente em corridas de galgos. Conforme detalhado na seção de Trabalhos Relacionados 2.6, uma busca sistemática revelou centenas de artigos sobre ML em esportes em geral, mas um número drasticamente reduzido, e muitas vezes nulo em algumas bases de dados, quando o foco se restringe a corridas de galgos. Essa escassez de pesquisas semelhantes, particularmente aquelas que visam a predição do tempo exato de corrida em vez de apenas o vencedor ou posições, justifica a relevância e a originalidade deste estudo.

A principal contribuição deste trabalho é justamente preencher essa lacuna, desenvolvendo um modelo preditivo para o tempo de corrida de galgos. Diferentemente dos poucos estudos existentes na área como (Lyons, 2016) ou (Schumaker; Johnson, 2008), que geralmente se concentram na previsão de resultados binários ou classificações de posição, a predição do tempo oferece uma métrica de desempenho mais detalhada. Essa abordagem granular permite inferências mais precisas sobre as posições finais e o desempenho relativo entre os competidores, possibilitando uma análise mais sofisticada das corridas.

Do ponto de vista tecnológico e prático, o presente estudo também se justifica pela exploração do *CatBoostRegressor*. Esta biblioteca de *Gradient Boosting* é particularmente adequada para lidar com as características dos dados de corrida de galgos, que frequentemente incluem variáveis categóricas de alta cardinalidade (como categoria da corrida e *trap* de largada). O *CatBoost* se destaca por seu tratamento inovador dessas variáveis e sua robustez contra *overfitting*, otimizando o processo de pré-processamento e a performance preditiva (Dorogush; Ershov; Gulin, 2018). Além disso, o trabalho aborda o desafio da qualidade e disponibilidade dos dados nesse

domínio, que muitas vezes são heterogêneos e exigem um esforço considerável em coleta, limpeza e padronização, demonstrando a construção de uma base de dados robusta a partir de informações disponíveis.

Em suma, a relevância econômica das corridas de galgos, aliada à ausência de ferramentas preditivas avançadas na literatura e no mercado, reforça a necessidade de abordagens mais objetivas e consistentes. Ao desenvolver um modelo de ML focado na predição de tempo e utilizando o potencial do *CatBoost*, este trabalho não só contribui para o avanço da pesquisa em inteligência artificial aplicada a esportes, como também oferece um benefício prático significativo para o mercado de apostas, superando as limitações da análise humana e dos sistemas estatísticos tradicionais.

1.2 Objetivos

Esta seção descreve o objetivo geral do trabalho e o detalha em objetivos específicos que são essenciais para sua execução.

1.2.1 Objetivo geral

Desenvolver um modelo de aprendizado de máquina utilizando o *CatBoostRegressor* para prever com precisão os tempos de corrida de galgos, a partir de dados históricos da temporada de 2024, visando auxiliar a aplicação de estratégias de apostas.

1.2.2 Objetivos específicos

- Analisar as regras e padrões das corridas de galgo do Reino Unido e definir as melhores condições de corridas para a seleção das corridas a serem utilizadas.
- Demonstrar o processo de definição e aplicação de uma *pipeline* estruturada de preparo de dados, avaliando a base de dados antes e após o processo.
- Demonstrar técnicas de otimização de hiperparâmetros para algoritmos de *machine learning* com diferentes métodos de avaliação.
- Apresentar análises dos resultados obtidos por meio do modelo e identificar estratégias de aposta dentre diferentes classificações disponíveis no mercado, calculando possíveis ganhos com as estratégias identificadas.

Para alcançar esses objetivos, é fundamental explorar o referencial teórico que fundamenta a aplicação de aprendizado de máquina nas corridas de galgos.

2 REFERENCIAL TEÓRICO

2.1 Corridas de Galgos no Reino Unido

Com origem no século XX, as corridas de galgos, importadas dos Estados Unidos em 1926 com a chegada da lebre mecânica, rapidamente se consolidaram como um dos esportes mais populares no Reino Unido. Inicialmente introduzidas comercialmente no Belle Vue Stadium, em Manchester, e expandindo-se para Liverpool e Londres. De acordo com (Huggins, 2007), a modalidade tornou-se, em meados da década de 1930, a terceira maior atividade comercial de lazer na Grã-Bretanha, superada apenas pelo cinema e pelo futebol. Esse sucesso se deveu em parte à sua acessibilidade, com muitas pistas localizadas em distritos urbanos densamente povoados e eventos ocorrendo à noite, após o expediente de trabalho, o que a popularizou como ”o hipódromo do homem pobre”, onde as apostas em dinheiro eram permitidas legalmente. Segundo (Roudaut, 2017), a popularidade das corridas de galgos impulsionou significativamente o setor de apostas, atraindo não apenas entusiastas, mas também um grande número de apostadores motivados pelo retorno financeiro. Contudo, a previsão de resultados nesse mercado apresenta desafios consideráveis devido à complexidade e à natureza volátil dos dados históricos empregados.

A corrida de galgos pode variar em distância, entre 270 e 714 metros. O formato oval da pista e a estrutura reposicionável das *traps* (gaiolas de largada) ao longo da pista, permitem definir qualquer distância para uma corrida. Em corridas mais longas, os galgos percorrem voltas completas até alcançarem a linha de chegada, estimulados por uma lebre robótica que se move ao longo da borda externa da pista, incentivando-os a manter a velocidade até cruzarem a linha de chegada. Uma postagem no site da pista (Oxford Stadium, 2024a) descreve como funcionam as divisões das categorias e as suas características, com a seguinte organização:

A Categoria A abrange as corridas médias, de 380 a 500 metros. Os galgos desta categoria são divididos em 3 grupos: a primeira divisão são os da A1 a A3, a segunda divisão da A4 a A6 e a terceira divisão da A7 a A11. Nessa lógica, quanto menor o número que acompanha a categoria, melhor ela se torna. Na Categoria HP, os galgos não largam na mesma posição, ou seja, partem de posições diferentes, alguns com vantagem, com o intuito de equilibrar a corrida, em que alguns galgos apresentam desempenho significativamente superior (Oxford Stadium, 2024a). A Categoria OR, do inglês Open Race, permite a participação de qualquer galgo independentemente da categoria. Envolve a elite dos galgos, muitos sendo de categorias A1 e A2, e também

galgos que disputam os grandes Derbys para a escolha do melhor galgo do ano. Já a Categoria D compreende as corridas curtas abaixo de 300 metros, e a numeração segue a mesma lógica da categoria A, mas com o escopo de D1 a D6. Para as corridas longas, a Categoria S abrange distâncias acima de 600 metros, com a numeração seguindo a mesma lógica da categoria A, mas com o escopo de S1 a S6. A Categoria B é destinada a corridas com galgos que nunca venceram uma corrida da categoria A, e os números seguem a mesma lógica da categoria A. Por fim, a Categoria H refere-se a corridas com barreiras, onde os galgos devem saltar (Oxford Stadium, 2024a).

Diversas métricas são coletadas ao longo do percurso, e são organizadas em *cards*, que são disponibilizados aos visitantes antes da corrida. No artigo (Oxford Stadium, 2024b) publicado por uma pista de corrida, descreve todos os dados disponibilizados, o *split*, por exemplo, mede o tempo que cada galgo leva para atingir a primeira curva após a largada oferece uma referência inicial de desempenho. Em cada curva subsequente, a posição de cada galgo é registrada, o que possibilita uma análise detalhada sobre seu comportamento durante a corrida e a possibilidade de identificar tendências, como se um galgo tende a ganhar ou perder posições ao longo do trajeto. Dado o cenário descrito, técnicas de aprendizado de máquina emergem como ferramentas promissoras para superar os desafios de análise preditiva em corridas de galgos

2.2 Aprendizado de Máquina e Previsão de Resultados

O uso de técnicas de Aprendizado de Máquina (ML) na previsão de resultados esportivos tem ganhado destaque devido ao crescimento dos mercados de apostas online e à vasta disponibilidade de dados históricos (Vaughan Williams; Stekler, 2010). Modelos de ML são capazes de identificar padrões complexos em grandes volumes de dados e adaptar suas previsões a diferentes combinações de variáveis, ajustando-se a tendências e comportamentos recorrentes de atletas e equipes (Zhu et al., 2024). Essa capacidade é crucial, especialmente em esportes com alta natureza estocástica, como o futebol, onde inúmeras características, desde habilidades individuais até moral da equipe e lesões, podem influenciar o resultado de uma partida (da Costa; Marinho; Pires, 2022).

Diversos estudos na literatura demonstram a aplicação bem-sucedida de ML em variadas modalidades esportivas. Na previsão de resultados em corridas de cavalos e galgos, por exemplo, algoritmos como Support Vector Regression (SVR) e Redes Neurais foram utilizados no trabalho (Schumaker, 2013) para prever posições de chegada e explorar ineficiências de mercado. Em esportes coletivos, como hóquei no gelo e futebol, sistemas de previsão baseados em ML empregaram análise de componentes

principais, testes estatísticos não paramétricos e métodos ensemble para alcançar alta precisão na previsão de vitórias e na identificação de fatores-chave de desempenho (Gu et al., 2019). Segundo (Fialho; Manhães; Teixeira, 2019), tais abordagens visam superar as limitações da análise manual, que é propensa a erros e enviesamentos humanos.

Apesar dos avanços, a eficácia dos modelos de ML na previsão esportiva está intrinsecamente ligada à qualidade e à engenharia das características utilizadas (Fialho; Manhães; Teixeira, 2019). A literatura enfatiza que problemas como valores ausentes, inconsistências e a complexidade dos dados históricos representam desafios significativos. De acordo com (da Costa; Marinho; Pires, 2022), a capacidade de extrair e modelar o conhecimento de domínio de forma eficaz, transformando dados brutos em atributos preditivos relevantes, é um fator determinante para o sucesso dos modelos de ML em cenários de previsão esportiva e na exploração de oportunidades em mercados de apostas.

2.3 *Gradient Boosting* e o Algoritmo *CatBoost*

O *Gradient Boosting* é uma técnica poderosa de aprendizado de máquina que alcança resultados de ponta em uma variedade de tarefas práticas. Por muitos anos, manteve-se como o método primário para problemas de aprendizado que envolvem *features* heterogêneas, dados ruidosos e dependências complexas (Hancock; Khoshgoftaar, 2020). Essencialmente, de acordo com (Natekin; Knoll, 2013), o *Gradient Boosting* constrói um preditor em ensemble realizando um gradiente descendente em um espaço funcional, combinando iterativamente modelos mais fracos para formar preditores mais fortes.

A flexibilidade do *Gradient Boosting Machines* (GBMs) permite alta customização para qualquer tarefa orientada a dados, oferecendo a escolha de diversas funções de perda e modelos base-learners, as implementações utilizam árvores de decisão como preditores base. Embora essas árvores de decisão sejam convenientes para *features* numéricas, muitos *datasets* na prática incluem *features* categóricas, que são também importantes para a previsão. A prática mais comum para lidar com *features* categóricas em *gradient boosting* é convertê-las para números antes do treinamento (Dorogush; Ershov; Gulin, 2018).

Nesse contexto, o *CatBoost* emerge como uma inovação significativa. O *CatBoost* é uma nova biblioteca de *gradient boosting* de código aberto que lida com *features* categóricas de forma bem-sucedida e supera implementações existentes de *gradient boosting* em termos de qualidade em diversos *datasets* públicos, conforme demonstrado por (Prokhorenkova et al., 2018). Duas de suas principais inovações algorítmicas

são a implementação do *ordered boosting*, uma alternativa baseada em permutações ao algoritmo clássico, e um algoritmo inovador para processar *features* categóricas. Ambas as técnicas foram desenvolvidas para combater um "desvio de previsão" causado por um tipo específico de *target leakage* presente nas implementações atuais de *gradient boosting*. O *CatBoost*, ao implementar essas modificações, evita o *target leakage* e aprimora a generalização do modelo (Prokhorenkova et al., 2018).

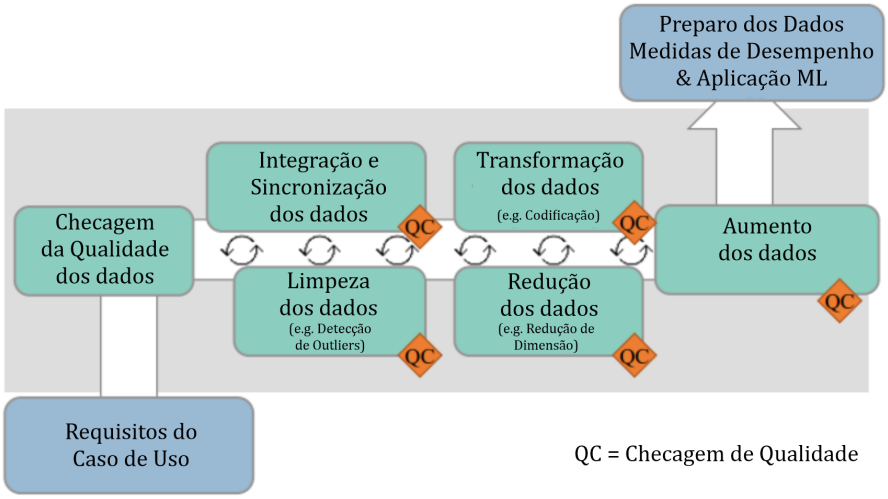
A implementação do *CatBoost* também se destaca pela sua performance e eficiência. Ele possui uma implementação em GPU para o algoritmo de aprendizado e uma implementação em CPU para o algoritmo de *scoring*, que são significativamente mais rápidas do que outras bibliotecas de *gradient boosting* em *ensembles* de tamanhos similares (Dorogush; Ershov; Gulin, 2018). Além disso, o *CatBoost* utiliza árvores de decisão *oblivious* como preditores base, que são equilibradas e menos propensas ao *overfitting*, permitindo acelerar a execução em tempo de teste. Sua capacidade de lidar com *features* categóricas diretamente durante o treinamento, sem a necessidade de pré-processamento manual como o *one-hot encoding* para alta cardinalidade, e de aplicar permutações aleatórias para estimar valores das folhas, contribui para sua eficácia e redução do *overfitting* (Prokhorenkova et al., 2018).

2.4 Estruturação de *Pipelines* de Tratamento e Preparação de Dados

Um dos principais passos para o treinamento de um modelo de ML é o preparo dos dados, que influencia diretamente nos resultados obtidos. Segundo o estudo (Frye; Schmitt, 2020), assegurar a qualidade de dados adequada é considerado um dos maiores desafios, a baixa qualidade de dados resulta em resultados de análise ruins, o que também é conhecido como princípio de entrada de lixo, saída de lixo GIGO.

Além disso, pesquisas recentes enfatizam o efeito prejudicial da baixa qualidade dos dados nos modelos de ML. (Jain et al., 2020) argumenta que a eficiência de um modelo de aprendizado de máquina é diretamente condicionada à qualidade dos dados utilizados. Problemas como valores ausentes e inconsistências podem levar a análises imprecisas. Dados rotulados incorretamente, por sua vez, geram decisões não confiáveis. Destacando a importância de métricas e transformações para preencher as falhas na qualidade dos dados (Jain et al., 2020). A Figura 1 representa os passos propostos no trabalho (Frye; Schmitt, 2020) para definição da estrutura da *pipeline* de tratamento de dados.

Figura 1 – Etapas principais da *pipeline* de preparação de dados.



Fonte: (Frye; Schmitt, 2020)

Outro estudo relevante, (Budach et al., 2022) de maneira empírica, demonstra como seis aspectos da qualidade dos dados influenciam o desempenho de modelos de classificação, regressão e agrupamento. Notou-se que até mesmo pequenas poluições nos dados podem comprometer consideravelmente a acurácia dos modelos.

Portanto, assegurar a qualidade dos dados não é apenas um passo crucial no processo de aprendizado de máquina, mas também uma prática essencial para prevenir decisões equivocadas e resultados enviesados, conforme destacado por diversos estudos no campo.

2.5 Validação Cruzada *K-Fold*

A validação cruzada *K-fold* é uma técnica fundamental na avaliação de modelos de aprendizado de máquina, amplamente utilizada para assegurar a confiabilidade e a capacidade de generalização de modelos preditivos. Essa metodologia consiste em dividir o conjunto de dados em k subconjuntos, ou ”*folds*”, onde o modelo é iterativamente treinado em k-1 *folds* e validado no *fold* restante (Nti et al., 2021). Este processo é repetido k vezes, garantindo que cada fold seja utilizado como conjunto de validação exatamente uma vez , o que proporciona uma estimativa mais robusta do desempenho do modelo em dados não vistos e ajuda a mitigar problemas como o *overfitting* e o viés de seleção (Bhagat; Bakariya, 2025). Além disso, a validação cruzada é essencial para o ajuste de hiperparâmetros e para a seleção do modelo, otimizando seu desempenho geral.

A escolha do valor de k na validação cruzada *K-fold* é um fator crítico que im-

pacta diretamente a performance do modelo e a complexidade computacional. Embora não exista uma regra formal, valores de k como 5 ou 10 são frequentemente utilizados na literatura, pois são considerados um bom equilíbrio entre viés e variância na estimativa da taxa de erro (Wong; Yeh, 2020). No entanto, estudos empíricos como o (Nti et al., 2021) indicam que o valor ótimo de k pode variar significativamente entre diferentes algoritmos de aprendizado de máquina e conjuntos de dados, não podendo ser generalizado para todas as situações. Pesquisas sugerem que valores menores de k resultam em menor custo computacional, mas maior viés, enquanto valores maiores de k são computacionalmente mais caros, mas têm menor viés e maior variância.

Segundo (Bhagat; Bakariya, 2025), validação cruzada *K-fold* permite obter uma estimativa mais robusta do desempenho de um modelo, comparada a métodos mais simples como a divisão única entre treino e teste. Ela é uma ferramenta poderosa para testar a taxa de sucesso de modelos, especialmente em tarefas de classificação de dados. A técnica é crucial para determinar quão bem um modelo se generalizará para dados não vistos e para identificar se há problemas de superajuste. Ao considerar a precisão e a capacidade de generalização, a validação cruzada se torna um método indispensável para a construção de modelos de ML confiáveis e eficazes.

2.6 Trabalhos Relacionados

A aplicação de técnicas de ML na previsão de resultados esportivos tem ganhado crescente atenção na literatura científica, abrangendo diversas modalidades. Contudo, ao analisar o panorama da pesquisa, observa-se uma notável disparidade na quantidade de estudos dedicados a diferentes esportes. Para contextualizar a relevância e a lacuna de pesquisa abordada neste trabalho, foi realizada uma busca sistemática em cinco bases de dados (ScienceDirect, IEEE Xplore, ACM Digital Library, Springer Nature e Google Scholar) utilizando palavras-chave relacionadas à aplicação de ML em esportes e, mais especificamente, em corridas de galgos. Os resultados desta pesquisa foram sumarizados nas Tabelas 1 e 2.

Tabela 1 – Busca por artigos aplicando ML em esportes

Palavras-chave	**ScienceDirect**	**IEEE Xplore**	**ACM**	**Springer Nature**	**Scholar**
”sports betting”AND ”machine learning”	50	4	3	77	200+
”sports prediction”AND ”machine learning”	31	12	0	17	200+

Fonte: Elaborada pelo autor

Tabela 2 – Busca por artigos aplicando ML em corridas de galgos

Palavras-chave	ScienceDirect	IEEE Xplore	ACM	Springer Nature	Scholar
”greyhound racing”AND ”machine learning”	4	3	0	0	29
”greyhound race”AND ”machine learning”	2	3	0	0	6

Fonte: Elaborada pelo autor

Os resultados revelam um cenário de abundância para a previsão esportiva em geral e uma escassez significativa de trabalhos focados em corridas de galgos. Por exemplo, a busca por “sports betting” AND “machine learning” e “sports prediction” AND “machine learning” retornou centenas de artigos no Google Scholar e dezenas em outras bases. Em contraste, as buscas por “greyhound racing” AND “machine learning” e “greyhound race” AND “machine learning” resultaram em um número drasticamente menor, com apenas 29 e 6 artigos, respectivamente, no Google Scholar, e em sua maioria zero ou um dígito nas demais bases. Essa disparidade quantitativa corrobora a existência de uma lacuna significativa na pesquisa sobre a aplicação de ML para previsão em corridas de galgos, justificando a relevância do presente estudo.

Dentre os poucos trabalhos identificados que abordam a aplicação de ML em contextos de corrida de animais, três se destacam pela sua proximidade temática e metodológica, merecendo uma análise mais aprofundada:

2.6.1 “An Investigation of SVM Regression to Predict Longshot Greyhound Races” por Robert P. Schumaker e James W. Johnson (2008)

Este trabalho seminal de Schumaker; Johnson é um dos poucos que abordam especificamente a previsão em corridas de galgos com o uso de aprendizado de máquina, com um foco particular em *longshots* (galgos com poucas chances de vitória, mas que oferecem grandes retornos em caso de acerto). O objetivo central foi investigar se a Regressão por *Support Vector Machine* (SVR) poderia ser utilizada para prever o sucesso de galgos azarões. Os autores empregaram 20 variáveis de entrada para o modelo, incluindo características do galgo (e.g., peso, velocidade média, vitórias), características da corrida (e.g., distância, tipo de pista), e informações sobre as cotas (*odds*) das apostas. A avaliação dos modelos foi realizada com base na acurácia de previsão e na lucratividade. Os resultados apresentaram uma acurácia de 78,9% para prever se um *longshot* venceria a corrida. Mais significativamente, o estudo demonstrou a potencial lucratividade da estratégia proposta, com retornos de 15,1% para apostas *Win* (vencedor), 18,3% para apostas *Place* (entre os dois primeiros) e impressionantes

64,8% para apostas *Show* (entre os três primeiros), ao focar em *longshots*. Embora este trabalho seja altamente relevante por focar em corridas de galgos e demonstrar o potencial de lucro, sua metodologia se concentra na classificação de *longshots* para apostas de posição, e não na previsão do tempo exato de corrida, que é o objetivo central da presente pesquisa, nem utiliza a abordagem de *Gradient Boosting* explorada aqui.

2.6.2 "Man v Machine: Greyhound Racing Predictions" por Alva Lyons (2016)

Lyons explora a capacidade de algoritmos de aprendizado de máquina para prever resultados em corridas de galgos, confrontando seu desempenho com a expertise humana. O principal objetivo do autor foi determinar se modelos de ML poderiam superar as previsões de apostadores humanos experientes. Para isso, o estudo utilizou dados históricos de corridas de galgos e empregou um processo de seleção de atributos algorítmico, o qual identificou 13 variáveis cruciais para a previsão. Estas variáveis incluíam atributos como a idade do galgo, peso, índice de massa corporal (IMC), posição da armadilha (*trap*), classificações de corrida anteriores (e.g., *grade*), e o número de vitórias/pódios recentes. A metodologia envolveu a comparação de modelos de ML com a performance de um apostador profissional. Os resultados indicaram que os algoritmos de ML, particularmente aqueles baseados em *Random Forest*, apresentaram uma taxa de acerto ligeiramente superior à do apostador humano em determinadas condições. No entanto, o estudo (Lyons, 2016) se concentrou primariamente na previsão do vencedor ou de posições finais, e não na previsão do tempo exato de corrida, o que representa uma diferença fundamental em relação ao objetivo deste trabalho. Além disso, o estudo não detalhou estratégias de apostas baseadas nos resultados preditos nem aprofundou na explicabilidade dos modelos.

2.6.3 "Machine learning in the prediction of flat horse racing results in Poland" por Piotr Borowski e Marcin Chlebus (2021)

Embora focado em corridas de cavalos e não em galgos, o estudo (Borowski; Chlebus et al., 2021) é pertinente por sua abordagem robusta no uso de ML para previsão de resultados de corrida, incluindo técnicas de explicabilidade (XAI). Os autores compararam a eficácia de seis algoritmos de classificação (CART, Glmnet, XGBoost, RF, NN e LDA) para criar estratégias de apostas lucrativas. Utilizando dados de corridas de cavalos na Polônia (2011-2020), aplicaram a técnica de *Variable Importance*, identificando o histórico de desempenho, posição de partida e distância da corrida como fatores cruciais. Os modelos de ML demonstraram capacidade de gerar lucro, com uma taxa de acerto de 41% para apostas *Win*. A relevância deste trabalho para

o presente estudo reside em sua metodologia sistemática de comparação de modelos e identificação de variáveis importantes, fornecendo um paralelo valioso apesar da diferença na modalidade esportiva.

2.6.4 Diferenciais do Presente Trabalho

Em comparação com os trabalhos relacionados, o presente trabalho propõe uma abordagem distinta e inovadora que visa preencher as lacunas identificadas. Primeiramente, diferente de (Lyons, 2016) e (Schumaker; Johnson, 2008), que se concentram na previsão de posições, este trabalho tem como objetivo principal prever o tempo exato de corrida de cada galgo. Essa abordagem permite uma análise mais granular do desempenho e pode ser utilizada para inferir posições de forma mais precisa, além de possibilitar estratégias de aposta baseadas em tempo esperado.

Em segundo lugar, enquanto os trabalhos anteriores utilizaram SVM, *Random Forest*, *XGBoost* e outras técnicas, este estudo explora o potencial do algoritmo *CatBoostRegressor*, um algoritmo de *Gradient Boosting* que possui otimizações para lidar com variáveis categóricas. Adicionalmente, o presente trabalho enfatiza a construção de uma *pipeline* robusta para a coleta, tratamento e engenharia de atributos a partir de dados brutos de corridas de galgos, uma etapa crucial e subdetalhada em outros estudos. Por fim, a previsão do tempo de corrida abre caminho para o desenvolvimento e avaliação de estratégias de apostas inovadoras que vão além das simples apostas de *Win* ou *Place*, permitindo a identificação de novas oportunidades no mercado.

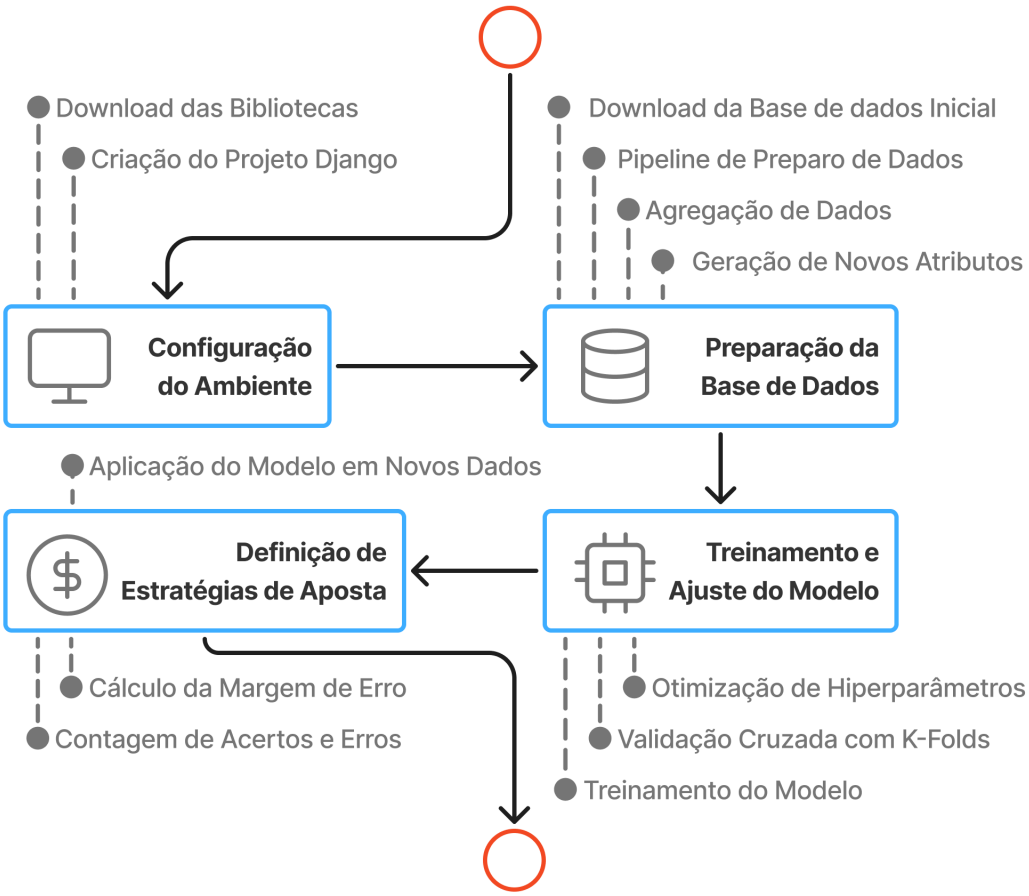
Em suma, embora a literatura demonstre a aplicabilidade do *Machine Learning* na previsão esportiva, a escassez de trabalhos específicos em corridas de galgos, aliada ao foco inovador na previsão de tempos de corrida e na exploração do *CatBoostRegressor*, posiciona este trabalho como uma contribuição relevante e original para o campo.

3 METODOLOGIA

Este capítulo descreve a metodologia utilizada para o desenvolvimento do estudo. Quanto à sua natureza, esta pesquisa caracteriza-se como quantitativa e aplicada, com um delineamento experimental. O foco metodológico reside na aplicação de um algoritmo de ML para a construção de um modelo preditivo, utilizando análise de dados e modelagem estatística para prever o tempo de corrida de galgos. Os procedimentos adotados seguiram uma pipeline de dados estruturada, incluindo a análise exploratória, o tratamento de dados, a engenharia de atributos e, por fim, a avaliação do modelo por meio de métricas de desempenho objetivas, como RMSE e MAE.

O processo foi estruturado em quatro etapas centrais: Configuração do Ambiente, Preparação da Base de Dados, Treinamento e Ajuste do Modelo e Definição de Estratégias de Aposta. Todo o fluxo de trabalho utilizado está descrito no fluxograma apresentado na Figura 2.

Figura 2 – Fluxograma da metodologia



Fonte: Elaborada pelo autor

3.1 Configuração do Ambiente

O ambiente de desenvolvimento e execução para este trabalho foi estabelecido em um sistema operacional Windows 11, utilizando a linguagem de programação *Python* na versão 3.12.6. O desenvolvimento do código e a gestão do projeto foram realizados no *Visual Studio Code* (VS Code), um ambiente de desenvolvimento integrado (IDE) que proporcionou um fluxo de trabalho eficiente para a codificação e depuração.

Para a manipulação e organização dos dados, a biblioteca *Pandas* foi empregada para a leitura de arquivos, estruturação em *DataFrames* e operações essenciais de pré-processamento. A biblioteca *NumPy* foi utilizada para dar suporte a operações numéricas de baixo nível, otimizando cálculos envolvendo *arrays* de dados.

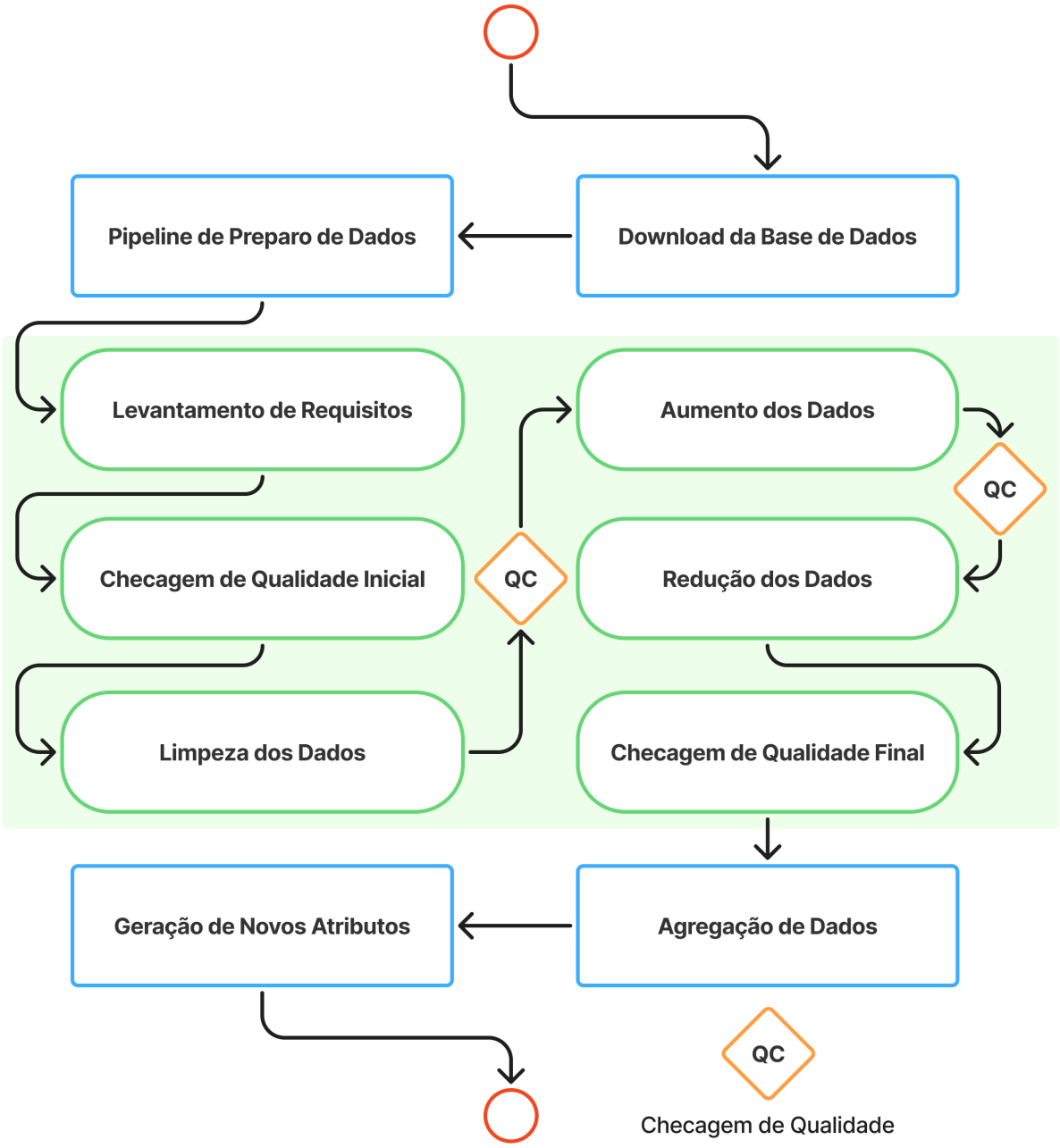
No que tange ao aprendizado de máquina, o *Scikit-learn* foi a base para diversas etapas, incluindo a divisão dos dados em conjuntos de treino e teste, a aplicação de técnicas de pré-processamento como padronização e codificação de variáveis, e a utilização de ferramentas para validação cruzada. O modelo preditivo principal foi construído com o *CatBoost*, especificamente o *CatBoostRegressor*, escolhido por sua capacidade de lidar de forma nativa com variáveis categóricas e sua robustez em modelos de *Gradient Boosting*.

Por fim, as bibliotecas *Matplotlib* e *Seaborn* foram utilizadas para a criação de visualizações gráficas. Essas ferramentas foram importantes para a exploração inicial dos dados e para a representação visual de características do conjunto de dados e dos resultados intermediários do processo. Com o ambiente de desenvolvimento devidamente configurado, a próxima etapa executada consiste no preparo da base de dados.

3.2 Preparação da Base de Dados

Os dados utilizados neste estudo consistem em registros públicos das corridas de galgos referentes à temporada de 2024. Ressalta-se que nenhum animal foi manuseado ou submetido a qualquer tipo de interação durante a realização da pesquisa, uma vez que todos os dados analisados são de domínio público e coletados de forma exclusivamente documental. A Figura 3 apresenta um fluxograma com os principais passos do preparo da base de dados.

Figura 3 – Fluxograma do preparo da base de dados



Fonte: Elaborada pelo autor

3.2.1 Dicionário de dados da base inicial

Inicialmente foi realizado o download da base de dados diretamente do site da (Greyhound Board of Great Britain, 2024). O Quadro 1 descreve cada atributo contido nos dados baixados.

Quadro 1: Dicionário de dados da base inicial

Campo	**Tipo**	**Descrição**
race_id	Inteiro	Número de identificação único da corrida.
track_id	Inteiro	Número de identificação único da pista.
dog_id	Inteiro	Número de identificação único do galgo.
date	Data	Data da corrida.
trap	Inteiro	Trap (armadilha ou caixa) de largada na corrida.
distance	Inteiro	Distância ou tamanho do percurso da corrida em metros.
split	Float	Tempo que o galgo leva da largada até a primeira curva.
bndPos	Texto	As posições do galgo em cada curva da pista.
finalPosition	Inteiro	Posição final do galgo.
trackShortName	Texto	Abreviação do nome da pista.
handicapMetre	Inteiro	Distância em metros de vantagem na largada do galgo (exclusivo em corridas HP).
winnersTimeS	Float	Tempo do galgo vencedor ou segundo colocado.
goingType	Inteiro	Condições da pista no momento da corrida.
grade	Texto	Categoria da corrida.
time	Float	Tempo no qual o galgo completou a corrida.
remarks	Texto	Comentários comportamentais do galgo durante a corrida.
weight	Float	Peso do galgo.

Fonte: Elaborado pelo autor

3.2.2 *Pipeline* de Preparo de Dados

A implementação seguiu uma metodologia fundamentada na literatura, contemplando as etapas destacadas na área verde no fluxograma apresentado na Figura 3. Foi empregada a biblioteca *YData Profiling* para a avaliação exploratória da base de dados.

Como primeiro passo, os requisitos foram determinados, de modo que a seleção dos métodos DPP foi altamente dependente dos requisitos definidos. Em termos gerais, foi determinado que os dados não podiam ter valores ausentes ou incorretos, nem valores discrepantes ou *outliers*. Para a variável alvo tempo, a ausência de valores incorretos foi considerada muito importante. Quanto à base de dados, foi estabelecido que não deveria haver dados duplicados, e que a base deveria ser limpa e concisa.

Em sequência, a pipeline iniciou-se com a checagem de qualidade inicial dos dados, onde inconsistências e valores ausentes foram identificados na base de dados

bruta. Seguindo o fluxo, procedeu-se à Limpeza dos Dados, por meio da aplicação de filtros e da remoção das observações com valores ausentes ou incorretos em atributos-chave, bem como a filtragem por características específicas de corridas. Foram removidos os os dados nos quais apresentaram valores ausentes ou zeros nos atributos *finalPosition*, *bndPos*, *split* e *time*. Foram filtradas as corridas com a categoria do tipo A ou OR , com a distância entre 460 e 500 metros. Após esta etapa, uma checagem de qualidade foi realizada para verificar a consistência da base de dados limpa.

Posteriormente, o processo avançou para o aumento dos dados, com a criação de novos atributos originados a partir de cálculos e relações de atributos já existentes. por meio desta etapa, foram criados os atributos descritos no Quadro 2, são informações importantes que descrevem ou representam a performance do galgo durante a corrida. Em seguida uma nova checagem de qualidade foi então executada para assegurar a integridade e a validade dos dados enriquecidos.

Quadro 2: Novos atributos

Campo	**Tipo**	**Descrição**
largada	Inteiro	Posição do galgo na primeira curva. extraído do atributo bndPos.
ultBend	Inteiro	Posição do galgo na última curva. extraído do atributo bndPos.
rec	Inteiro	Posições ganhas ou perdidas durante a corrida. Resultante de (largada - finalPosition).
splitFin	Float	Posições ganhas ou perdidas na reta final da corrida. Resultante de (ultBend - finalPosition).
velMed	Float	Velocidade média do galgo. Resultante de (*distance* / *time*).

Fonte: Elaborado pelo autor

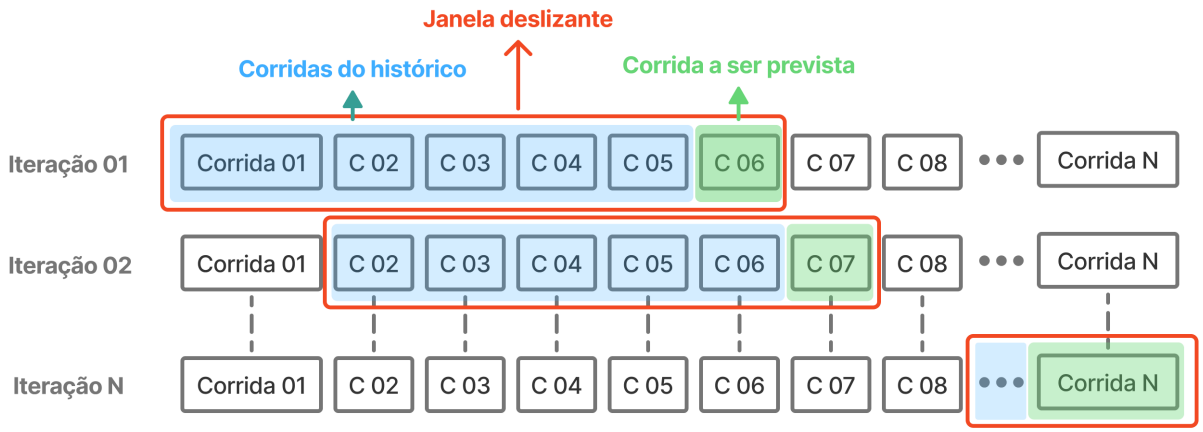
A checagem possibilitou a identificação de atributos que não contribuem para a previsão, e para atingir os requisitos da base de dados, foi aplicada a redução dos dados, realizada por meio da remoção de atributos. Nesta etapa foram removidos 11 atributos, dentre eles, identificadores únicos, comentários textuais da corrida, atributos com alta taxa de ausência, atributos que se tornaram redundantes após a criação de novos e atributos com baixa interação com a variável alvo *time*. Foi o último passo de processamento principal da pipeline, antes da uma checagem final da qualidade dos dados. A checagem de qualidade final dos dados representou o encerramento da *pipeline* de preparo de dados, que validou a conformidade do conjunto de dados para os próximos processos de tratamento.

3.2.3 Agregação de Dados e Geração de Novos Atributos

Após a conclusão da *pipeline* de preparos de dados, o conjunto de dados foi processado para gerar novos atributos e agregar informações relevantes para cada galgo, formando a base final para o treinamento do modelo de regressão.

Os dados históricos de corridas foram agrupados pela distância da corrida para cada galgo distinto. Dentro de cada grupo, as corridas foram ordenadas cronologicamente. Em seguida, foi aplicada uma metodologia de janela deslizante apresentada na Figura 4 para calcular médias e outras métricas estatísticas baseadas nos resultados históricos de cada galgo. Isso significa que, para cada corrida que um galgo competiu, as características preditivas foram derivadas das 5 corridas anteriores daquele mesmo galgo na mesma distância. Foram aceitos apenas dados em que o galgo possui pelo menos 6 corridas oficiais naquela determinada distância (5 para histórico + 1 a ser prevista).

Figura 4 – Fluxograma janela deslizante



Fonte: Elaborada pelo autor

A partir desse processo, foi gerado um novo *dataset* com os atributos descritos no Quadro 3 para compor a base de dados final. O novo *dataset* reuniu as principais características de desempenho do galgo, como a variação média de tempo entre as corridas, o valor representa a constância do galgo na sua carreira. O atributo *overall*, consiste no coeficiente de tendência linear dos tempos calculado por uma regressão linear simples e apresenta um valor numérico do quanto o galgo melhorou ou piorou seu desempenho nas ultimas corridas.

Quadro 3: Base de dados final

Atributo	**Tipo**	**Descrição**
timeMed	Float	Média de Tempo.
finMed	Float	Média de finalização.
posLargMed	Float	Posição de largada média.
splitFinMed	Float	Média de split final.
recMed	Float	Média de recuperação.
velMed	Float	Velocidade média do galgo.
varMed	Float	Variância média dos tempos anteriores do galgo.
overall	Float	Coeficiente de tendência linear dos tempos históricos do galgo.
trap	Texto	Trap (armadilha ou caixa) de largada na corrida a ser prevista.
modTrap	Texto	Trap de costume do galgo.
distance	Inteiro	Distância da corrida a ser prevista.
raceCat	Texto	Categoria da corrida a ser prevista.
ultCat	Texto	Categoria da última corrida do galgo.
time	Float	Tempo realizado pelo galgo na corrida a ser prevista.

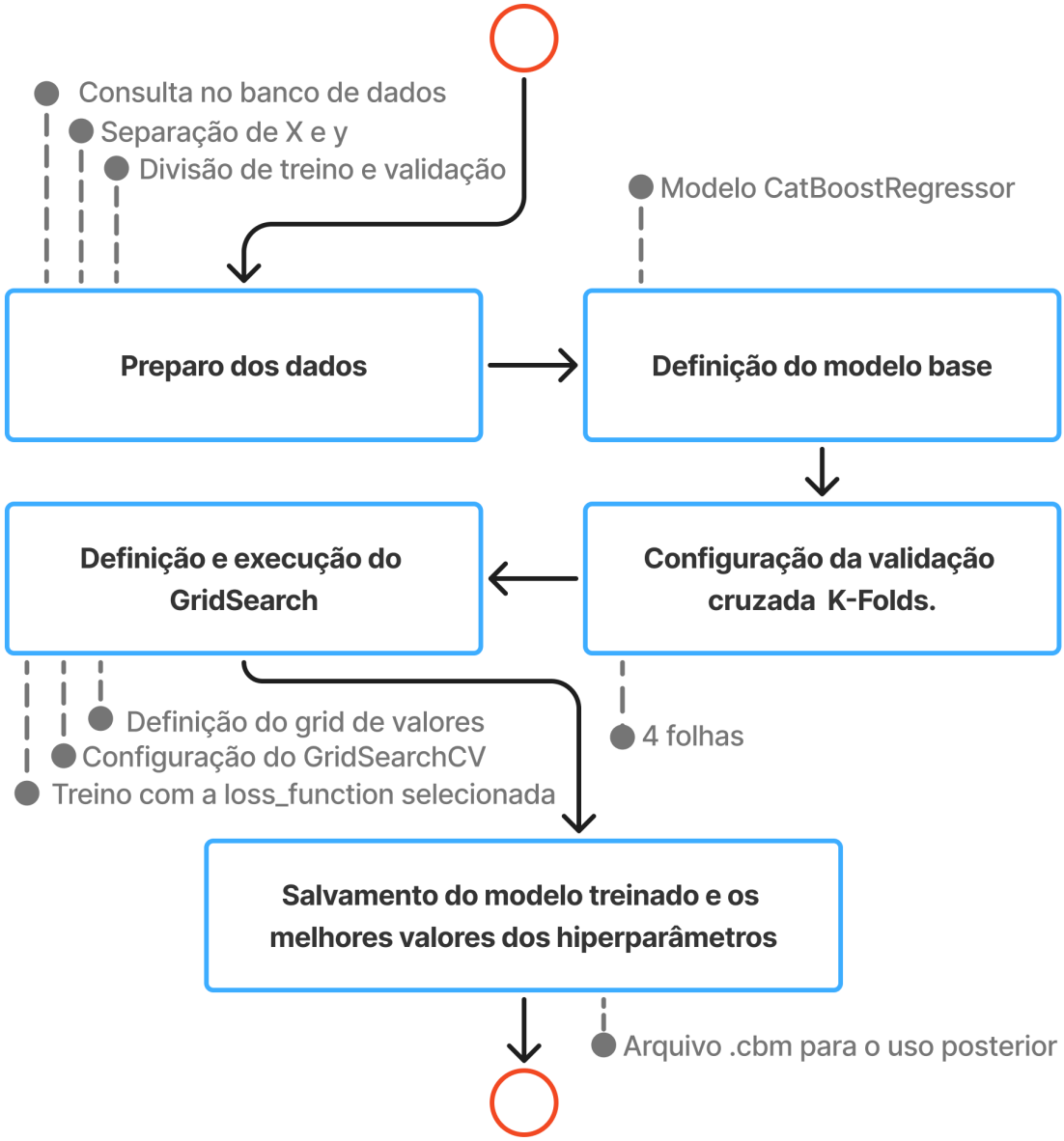
Fonte: Elaborado pelo autor

Após a criação do *dataset* final, foi realizada uma checagem de qualidade que seguiu os mesmos princípios da *pipeline* de tratamento de dados, confirmando sua adequação e consistência para as fases subsequentes de treinamento e ajuste do modelo.

3.3 Treinamento e Ajuste do Modelo

O processo de treinamento e otimização do modelo de regressão *CatBoostRegressor* foi conduzido para encontrar a melhor configuração de hiperparâmetros para a predição de tempo. Este processo envolveu diversas etapas cruciais, desde a preparação dos dados até a avaliação final do modelo. A função implementada para realizar este processo está disponível no Apêndice A.1. Esta metodologia foi aplicada de forma independente para as funções de perda Root Mean Squared Error (RMSE) e Mean Absolute Error (MAE). A ilustração detalhada deste processo é apresentada na Figura 5.

Figura 5 – Fluxograma do treinamento e ajuste do modelo



Fonte: Elaborada pelo autor

O desenvolvimento do modelo de regressão para predição do tempo de corrida envolveu um processo iterativo de treinamento e otimização, fundamentado na utilização da biblioteca *CatBoost* e na técnica de busca exaustiva de hiperparâmetros por meio do *GridSearchCV*. As etapas principais deste processo são detalhadas a seguir. Inicialmente, os dados cruciais para o treinamento foram extraídos do banco de dados, selecionando os atributos definidos na etapa anterior, juntamente com a variável alvo *time*. Estes dados foram posteriormente estruturados em um *DataFrame Pandas*, um formato essencial para a manipulação e processamento eficiente pelas bibliotecas de aprendizado de máquina. Uma etapa preparatória fundamental consistiu na separação do conjunto de dados em duas partes distintas: o conjunto de treinamento, correspon-

dendo a 80% dos dados, e o conjunto de validação, compreendendo os 20% restantes. Esta divisão, realizada com uma semente aleatória específica (random_state=42), garantiu a consistência e a reprodutibilidade da separação, permitindo que o modelo fosse treinado e ajustado nos dados de treinamento e, posteriormente, avaliado em um conjunto de dados não visto, fornecendo uma estimativa da sua capacidade de generalização.

A arquitetura base do modelo foi definida pela instanciação de um regressor *CatBoost* o *CatBoostRegressor*, com a definição da mesma semente aleatória para assegurar a repetibilidade dos experimentos. A função de perda e a métrica de avaliação foram ambas referenciadas pela variável *loss_function* fornecida. Para uma avaliação mais aprofundada do desempenho do modelo durante a fase de ajuste de hiperparâmetros, implementou-se a técnica de validação cruzada *K-Fold*, na linha 26 do codigo fonte A.1 com um total de quatro *folds*. Esta abordagem envolveu a divisão do conjunto de treinamento em quatro partes, onde cada uma delas serviu como conjunto de teste em uma iteração, enquanto as demais foram utilizadas para o treinamento, com os dados sendo embaralhados e a semente aleatória definida para garantir a consistência do processo.

Para identificar a configuração ideal de hiperparâmetros, explorou-se um espaço de busca predefinido, cujas combinações detalhadas podem ser consultadas no Quadro 4. A busca sistemática por meio dessas combinações foi realizada utilizando o *GridSearchCV* da biblioteca *Scikit-learn*. Este método empregou o modelo *CatBoost* como modelo base e a estratégia de validação cruzada *K-Fold* definida anteriormente. A busca foi configurada para utilizar dois núcleos de processamento em paralelo.

Quadro 4: Hiperparâmetros e seus valores no *GridSearch*

Hiperparâmetro	**Descrição**	**Valores Testados**
iterations	Número de árvores de decisão (epochs de boosting).	[200, 500, 1000]
learning_rate	Taxa de aprendizado do modelo.	[0.01, 0.03, 0.1]
depth	Profundidade máxima das árvores de decisão.	[4, 6, 8]
l2_leaf_reg	Coeficiente de regularização L2.	[2, 5, 9, 11]
border_count	Número máximo de bins para atributos numéricos.	[32, 64, 128]

Fonte: Elaborado pelo autor

A execução da busca em grade foi iniciada, com uma atenção especial ao tratamento das características categóricas identificadas, as quais foram especificadas através do parâmetro *cat_features* no dicionário *fit_params_for_catboost*, definido na linha 43 do código fonte A.1. Adicionalmente, para otimizar o tempo de treinamento

e mitigar o risco de *overfitting*, implementou-se um mecanismo de parada antecipada. Este critério interrompeu o treinamento caso não houvesse melhoria no desempenho do modelo no conjunto de validação por cem iterações consecutivas. Ao término da busca, os melhores hiperparâmetros identificados e a correspondente melhor pontuação média obtida durante a validação cruzada foram registrados. O modelo treinado com esta combinação ótima de hiperparâmetros foi então avaliado no conjunto de validação reservado, através do cálculo do RMSE entre as previsões e os valores reais. Finalmente, o modelo com o melhor desempenho foi armazenado em um arquivo .cbm, facilitando sua utilização nas etapas subsequentes do projeto.

3.4 Definição de Estratégias de Aposta

Para avaliar o modelo em um contexto real de aposta, ele foi aplicado em dados de corridas nunca vistos. Foram selecionadas 1131 corridas realizadas em 2025 com as mesmas características das utilizadas no treinamento do modelo. Para cada corrida, realizou-se a previsão de tempo para cada galgo participante e, em seguida, projetou-se uma classificação, ordenando os tempos de modo crescente.

Para aprimorar a precisão das classificações do modelo nessas corridas, foi empregada uma estratégia de margem de segurança baseada no tempo previsto. Essa margem foi calculada para três tipos de classificação: na *Win*, que consiste no galgo vencer a corrida, a margem foi determinada pela diferença de tempo para o segundo colocado. Na *Top 3*, que consiste no galgo finalizar dentro das três primeiras posições, a margem foi calculada para cada galgo incluído nesta classificação, determinada pela diferença de tempo para o quarto colocado. E para a *Lay Placed*, que caracteriza uma aposta contra, na qual o galgo não pode alcançar a segunda posição, a margem foi calculada para cada galgo da terceira à sexta colocação, determinada pela diferença de tempo para o segundo colocado. Estas margens foram salvas em *DataFrames Pandas* para avaliações posteriores.

Em seguida, utilizou-se uma busca em grade para identificar os valores ótimos dessas margens. Os valores testados variaram de 0,05 a 0,50 segundos, com um incremento de 0,01 segundos em cada iteração. Com os melhores valores para a margem de segurança definidos, realizaram-se os cálculos de acertividade, considerando os melhores valores para cada classificação descrita anteriormente. Com os passos finais da metodologia concluídos, o próximo capítulo apresentará uma análise abrangente dos resultados obtidos em todas as etapas, fornecendo uma visão completa do desempenho do modelo.

4 RESULTADOS

Este capítulo apresenta a análise dos resultados alcançados pela metodologia empregada na pesquisa, desde o preparo dos dados até a avaliação de desempenho do modelo treinado. Os resultados obtidos demonstram o alcance de cada um dos objetivos específicos propostos, consolidando o desenvolvimento do modelo preditivo para tempos de corrida de galgos.

4.1 Análise dos Objetivos Específicos

Para que fosse possível realizar a análise das regras e padrões das corridas de galgos do Reino Unido e definir as melhores condições para a seleção das corridas, foi identificada a necessidade de filtrar os dados brutos focando em categorias específicas de corrida (tipo ’A’ ou ’OR’) e distâncias (entre 460 e 500 metros). Essa seleção permitiu a concentração do estudo em um subconjunto de 153.735 observações, representando aproximadamente 38% do volume inicial de dados, garantindo a relevância e a homogeneidade das informações para o treinamento do modelo.

O processo de definição e aplicação de uma pipeline estruturada de preparo de dados foi efetivamente demonstrado, ao se obter uma melhoria considerável na qualidade da base de dados. Isso foi evidenciado pela remoção de valores ausentes em atributos-chave como *finalPosition* (6,8% removidos) , *bndPos* (11,8% removidos) , e split (29,3% removidos). Além disso, 2,5% de registros com valor zero no atributo time foram tratados , resultando em um conjunto de dados final sem células ausentes ou linhas duplicadas. A pipeline também incluiu a criação de novos atributos importantes, como largada, *ultBend*, *rec*, *splitFin*, e *velMed*, que enriqueceram a base para a predição.

As técnicas de otimização de hiperparâmetros para o algoritmo *CatBoostRegressor* foram demonstradas com sucesso, com a aplicação de métodos de avaliação como a validação cruzada *K-Fold*. Através de um *GridSearchCV* que explorou combinações de hiperparâmetros como *iterations* ([200, 500, 1000]), *learning_rate* ([0.01, 0.03, 0.1]), e *depth* ([4, 6, 8]), foi possível aprimorar o desempenho do modelo preditivo. Os melhores valores obtidos para o modelo foram RMSE de 0.31059 e MAE de 0.23769 no conjunto de validação , com os hiperparâmetros selecionados de *iterations*=500, *learning_rate*=0.03, e *depth*=8.

Por fim, a análise dos resultados e a identificação de estratégias de aposta foram realizadas com êxito em um conjunto de 1.131 corridas nunca vistas de 2025.

Para a estratégia *Win*, foi obtida uma precisão de 42,11% com uma margem de 0,15 segundos em 76 corridas elegíveis. Na estratégia *Top 3*, a margem de 0,12 segundos resultou em uma taxa de acerto de 75,21% em 351 apostas aptas. Para *Lay Placed*, uma margem de 0,36 segundos alcançou uma taxa de acerto de 82,2% em 62 apostas aptas. Contudo, o cálculo dos ganhos potenciais foi inviabilizado pela indisponibilidade de cotações de retorno, devido a novas leis esportivas (Lei n.º 14.790/2023 e Portaria n.º 125/2024 do Ministério do Esporte).

4.2 Análises na Preparação da Base de Dados

Na checagem inicial de qualidade dos dados, a Tabela 3 apresenta as estatísticas da base de dados inicial, com mais de 400 mil observações e 17 variáveis, descritas anteriormente no Quadro 1.

Tabela 3 – Estatísticas do conjunto de dados inicial

Característica	**Quantidade**
Número de variáveis	17
Número de observações	400277
Células ausentes	587912
Linhas duplicadas	0

Fonte: Elaborada pelo autor

Em sequência, no processo de limpeza da base de dados, o atributo *finalPosition* apresentou 6,8% de valores ausentes, conforme a Tabela 4. Estes dados foram removidos para garantir dados completos sobre a colocação dos galgos, dada sua relevância direta e baixa proporção de ausência.

Tabela 4 – Estatísticas do atributo *finalPosition*

Característica	**Quantidade**
Ausente	27218
Infinito	0
Zeros	0
Negativo	0

Fonte: Elaborada pelo autor

Do mesmo modo, o atributo bndPos (11,8% ausente, Tabela 5), essencial para a performance intermediária, teve suas observações com valores faltantes removidas para manter a integridade dos dados de trajetória. De forma similar, o split apresentou

29,3% de dados ausentes, como constatado na Figura 6, este atributo o tempo até a primeira curva, também foi constatado valores muito extremos, de 0,01 a 84. De tal modo, também teve suas observações com *outliers* ou faltantes removidas.

Tabela 5 – Estatísticas do atributo *bndPos*

Valor	**Contagem**	**Frequência (%)**
Ausente	47077	11,8%
1111	27605	6,9%
2222	12479	3,4%

Fonte: Elaborada pelo autor

Tabela 6 – Estatísticas do atributo *split*

Característica	**Quantidade**
Ausente	117309
Infinito	0
Zeros	0
Negativo	0
Mínimo	0,01
Máximo	84

Fonte: Elaborada pelo autor

Adicionalmente, foram tratadas inconsistências no atributo de tempo. O atributo *time*, que registra o tempo total da corrida do galgo, não possuía valores ausentes, mas apresentou 2,5% de registros com o valor igual a zero (Tabela 7). Estes valores foram classificados como inconsistências de registro que que invalidavam a validade da observação. Para tratar esta questão e garantir a validade dos dados de tempo, todas as observações onde time era igual a zero foram removidas do dataset.

Tabela 7 – Estatísticas do atributo *time*

Característica	**Quantidade**
Ausente	0
Infinito	0
Zeros	9936
Negativo	0
Mínimo	0
Máximo	71,89

Fonte: Elaborada pelo autor

Para otimizar o modelo em condições de corrida mais representativas, foram aplicados filtros seletivos. Corridas entre 460 e 500 metros foram selecionadas pela predominância e homogeneidade na dinâmica (Tabela 8), visando otimizar o processo de aprendizagem.

Tabela 8 – Estatísticas do atributo *grade*

Valor	**Contagem**	**Frequência (%)**
A4	34777	8,7%
A5	33031	8,3%
A3	32697	8,2%
A6	27253	6,8%
A2	25070	6,3%
OR	23610	5,9%
A7	23439	5,9%
Outros Valores	200211	50,1%

Fonte: Elaborada pelo autor

Além disso, as categorias de corrida do tipo 'A' (*races* de alta qualidade e competitividade) e 'OR' (*Open Races*) somaram 50,1% das observações (Tabela 9). A escolha por essas categorias visa concentrar o estudo nas corridas de maior relevância e no tipo de evento que é mais frequentemente objeto de apostas e análises aprofundadas.

Tabela 9 – Estatísticas do atributo *distance*

Valor	**Contagem**	**Frequência (%)**
480	116796	29,2%
500	47249	11,8%
450	31952	8,0%
462	20276	5,1%
400	17511	4,4%
Outros Valores	166493	41,7%

Fonte: Elaborada pelo autor

Após a aplicação desses procedimentos de limpeza e filtragem, a base de dados foi significativamente refinada, removendo ruídos e inconsistências e focando nas observações mais relevantes e de maior qualidade para o objetivo do projeto. O volume final de dados e o impacto detalhado desta etapa na distribuição das características estão representados na Tabela 10, com 153735 observações, aproximadamente 38% do volume inicial.

Tabela 10 – Estatísticas do conjunto de dados após a limpeza

Característica	**Quantidade**
Número de variáveis	17
Número de observações	153735
Células ausentes	154183
Linhas duplicadas	0

Fonte: Elaborada pelo autor

A redução de dados foi uma etapa crucial para otimizar o modelo, eliminando atributos que não contribuem para a previsão e podem introduzir ruído nos dados, afetando a precisão. Atributos foram removidos com base na sua irrelevância, redundância ou falta de dados.

O campo *remarks* foi excluído por sua complexidade textual e alta variabilidade, evitando ruído e excesso de dimensionalidade. Identificadores únicos (race_id, dog_id, track_id) foram descartados por não se relacionarem com as previsões. O atributo *bndpos* foi removido devido à redundância, pois suas informações já foram extraídas e incorporadas. Por fim, *handicapMetre* foi excluído por apresentar 100% de valores nulos, e é exclusivo de categorias de corrida não analisadas (Tabela 11).

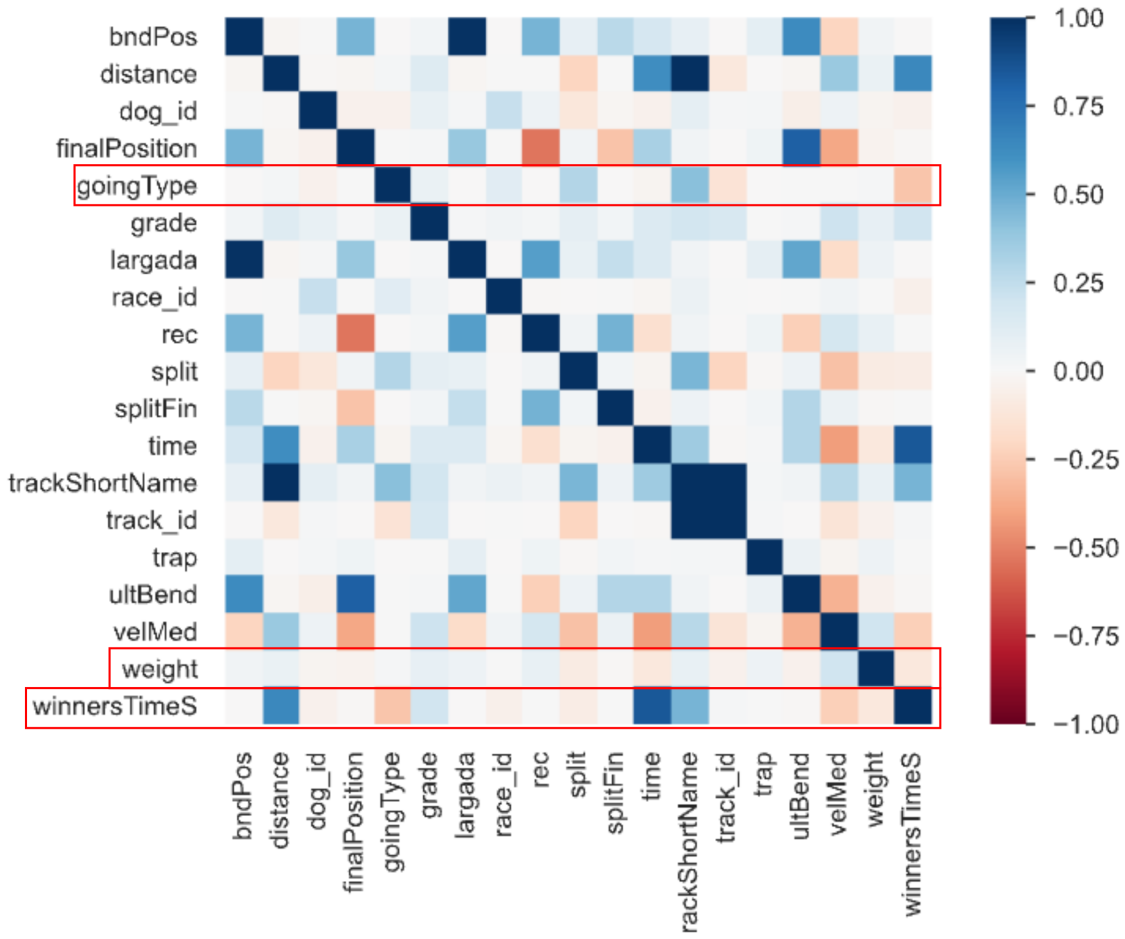
Tabela 11 – Estatísticas do atributo *handicapMetre*

Característica	**Quantidade**
Ausente	153735

Fonte: Elaborada pelo autor

Adicionalmente, os campos *winnersTimeS*, *goingType* e *weight* foram removidos do conjunto de dados devido à sua baixa influência e interação com os demais atributos, conforme evidenciado no mapa de calor de correlações da Figura 6. A fraca correlação com a variável alvo e com outras características importantes indicou que a manutenção desses atributos adicionaria pouca ou nenhuma informação relevante para o modelo preditivo, potencialmente introduzindo ruído.

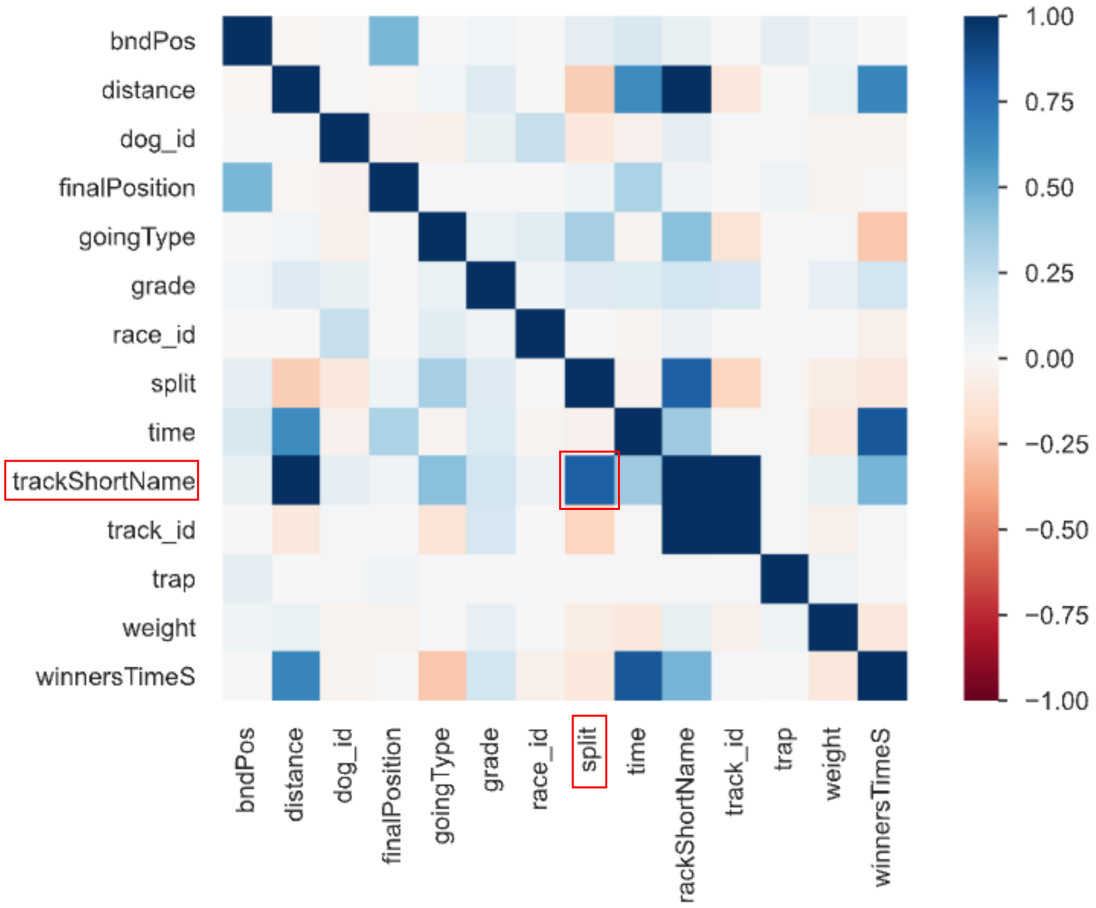
Figura 6 – Mapa de calor de correlações



Fonte: Elaborada pelo autor

Por fim, os atributos *split* e *trackShortName* também foram excluídos devido à sua alta correlação, conforme demonstrado na Figura 7. Essa forte dependência sugere que o split está mais ligado à pista do que à velocidade do galgo, e a permanência de ambos geraria redundância e multicolinearidade, potencialmente introduzindo ruídos no modelo.

Figura 7 – Mapa de calor de correlação *split trackShortName*



Fonte: Elaborada pelo autor

Finalmente, como último passo da pipeline de preparação de dados, a Tabela 12 ilustra a checagem final de qualidade e consistência do *dataset*, confirmando sua adequação para o treinamento do modelo.

Tabela 12 – Estatísticas do conjunto de dados final

Característica	**Quantidade**
Número de variáveis	11
Número de observações	153735
Células ausentes	0
Linhas duplicadas	0

Fonte: Elaborada pelo autor

4.3 Avaliação de Desempenho no Treinamento e Otimização

Os resultados da avaliação final do modelo no conjunto de validação, sumarizados no Quadro 13, fornecem uma medida direta do desempenho preditivo do CatBoostRegressor em dados não vistos. Para ambas as funções de perda (RMSE e MAE) utilizadas na avaliação, o modelo obteve o melhor desempenho com os mesmos valores dos hiperparâmetros. O Quadro 5 descreve os valores selecionados.

Tabela 13 – Resultados por função de perda

Função	**Score**	**Desvio med**	**Desvio max**	**Desvio min**	***Outliers***
RMSE	0.31059	0.0021	2.4243	-1.6028	207 (0,0096%)
MAE	0.23769	0.0276	2.4349	-1.5771	227 (0,0105%)

Fonte: Elaborado pelo autor

Quadro 5: Valores dos hiperparâmetros selecionados

Hiperparâmetro	**Valor Selecionado**
iterations	500
learning_rate	0.03
depth	8
l2_leaf_reg	9
border_count	128

Fonte: Elaborado pelo autor

O modelo otimizado com a função de perda RMSE apresentou um score de 0.31059 e um desvio médio de 0.0021. O modelo otimizado com a função de perda MAE, por sua vez, registrou um score de 0.23769 e um desvio médio de 0.0276. Apesar de ambos os modelos terem exibido baixa porcentagem de *outliers*, definidos como valores que excedem três desvios padrão da média dos erros, o desempenho do modelo otimizado por RMSE foi considerado superior, pois obteve um desvio médio significativamente menor, indicando maior consistência nas previsões em dados não vistos.

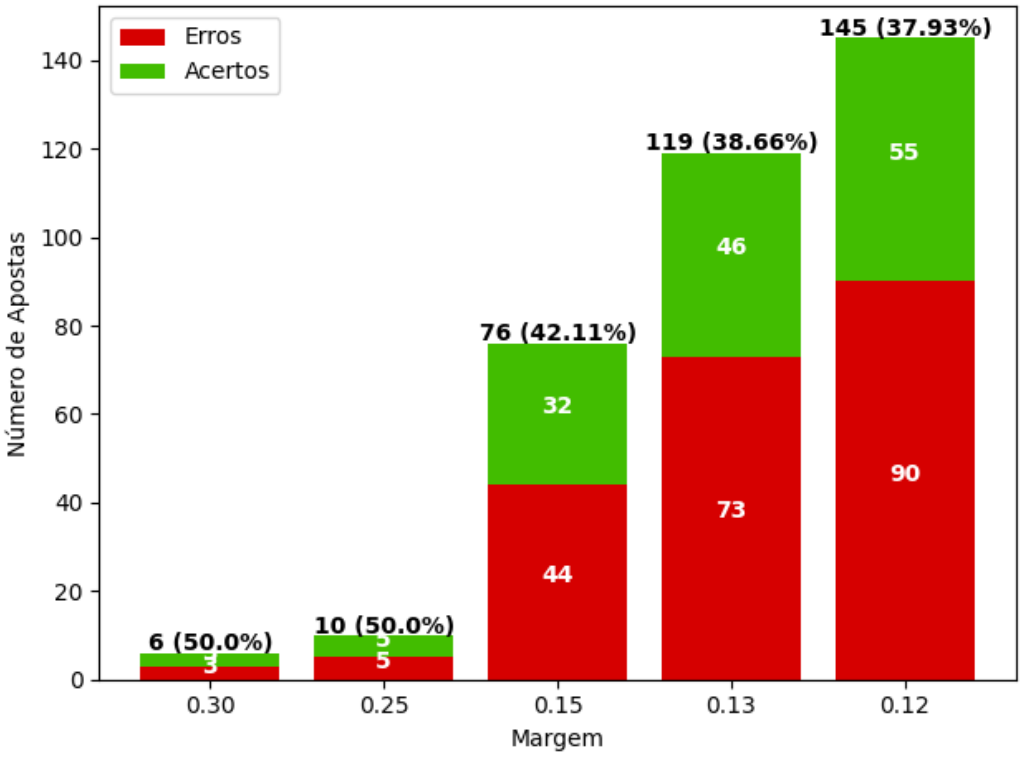
4.4 Desempenho nas Estratégias de Aposta

A seguir, apresenta-se o detalhamento dos resultados obtidos de cada estratégia de aposta definida na metodologia no item 3.4. Contudo, a estimativa de ganhos financeiros foi inviabilizada por novas regulamentações nacionais. A Lei nº 14.790, de 29 de dezembro de 2023 (Brasil, 2023), e a Portaria nº 125, de 31 de dezembro de

2024 (Ministério do Esporte, 2024), do Ministério do Esporte, regulamentaram as modalidades esportivas para apostas de quota fixa. Como as corridas de galgos não foram incluídas nessas modalidades autorizadas, as cotações de retorno ficaram indisponíveis no mercado regulamentado, impedindo a projeção de lucros.

Na estratégia *Win*, a previsão inicial do modelo sem a aplicação da margem de segurança resultou em 332 acertos em 1131 corridas, representando uma taxa de sucesso de 29,35%. A busca em grade subsequente, para otimizar o valor da margem de segurança, identificou os cinco melhores desempenhos, conforme detalhado na Figura 8. Buscando um equilíbrio entre a quantidade de apostas consideradas viáveis e o percentual de acertos, o valor de 0,15 segundos para a margem se destacou, alcançando uma precisão de 42,11% com 76 corridas elegíveis para aposta.

Figura 8 – Acertos vs Erros - *Win*

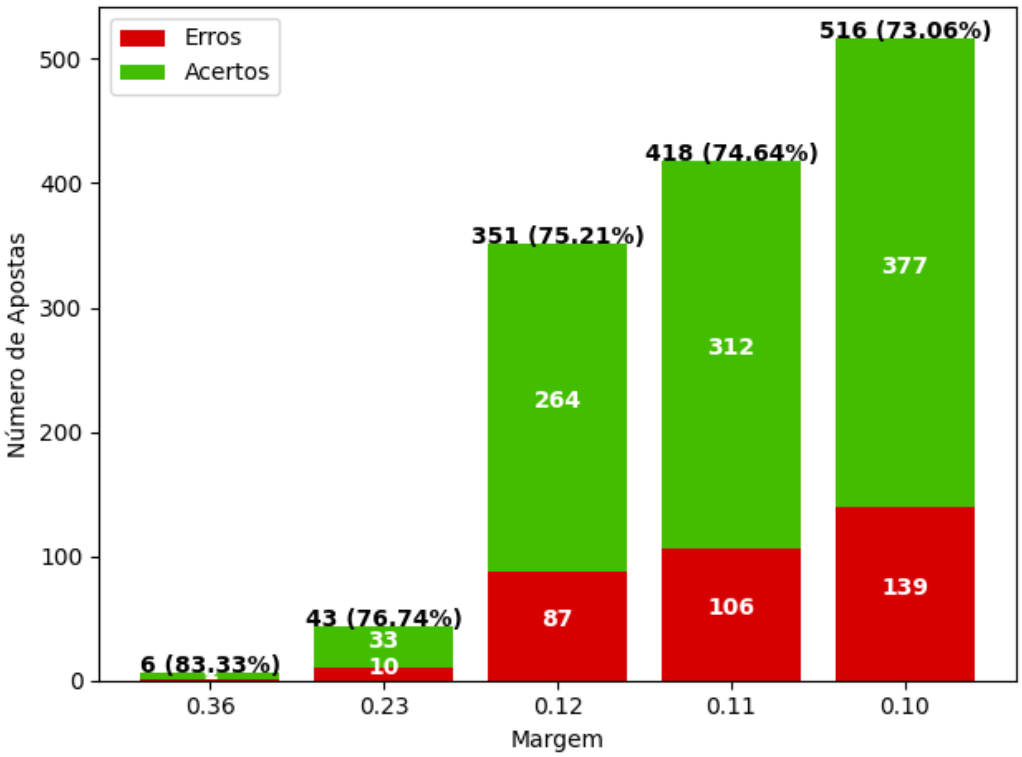


Fonte: Elaborada pelo autor

Para a estratégia *Top 3*, o modelo demonstrou uma assertividade inicial de 65,16% ao classificar corretamente 2211 galgos para o pódio, de um total de 3393. Similarmente à análise da estratégia *Win*, uma busca em grade foi realizada para determinar a margem de segurança ideal, cujos cinco melhores valores e respectivos resultados são apresentados na Figura 9. A análise indicou que uma margem de 0,12 segundos se mostrou a mais eficaz, proporcionando uma taxa de acerto de 75,21%

com 351 apostas aptas, mantendo um bom compromisso entre precisão e oportunidades de aposta.

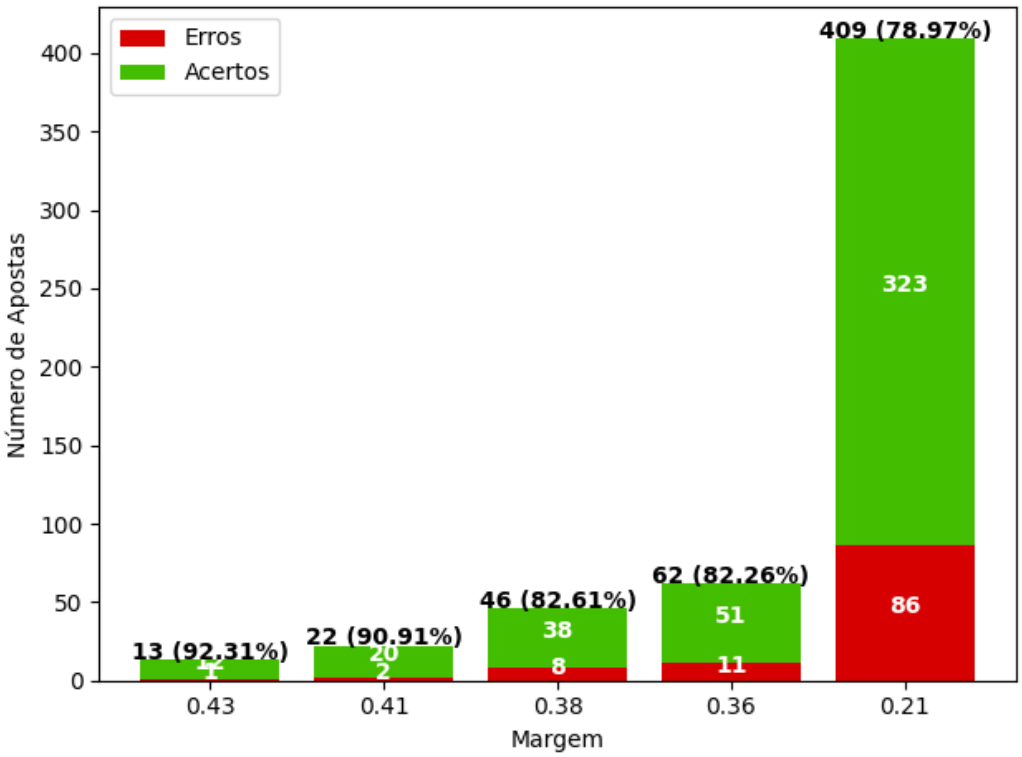
Figura 9 – Acertos vs Erros - Top 3



Fonte: Elaborada pelo autor

Na estratégia *Lay Placed*, a capacidade do modelo de prever corretamente a não classificação de galgos alcançou 67,92%, com 2479 previsões corretas em 3650. A otimização da margem de segurança por meio da busca em grade, cujos cinco principais resultados estão ilustrados na Figura 10, revelou que o valor de 0,36 segundos otimizou o desempenho. Esta configuração resultou em uma notável taxa de acerto de 82,2% com 62 apostas consideradas aptas, demonstrando uma alta confiança nas previsões realizadas sob essa margem.

Figura 10 – Acertos vs Erros - Lay



Fonte: Elaborada pelo autor

5 CONCLUSÕES

Este trabalho de conclusão de curso atingiu seu objetivo geral de desenvolver um modelo de aprendizado de máquina para a previsão de tempos de corrida de galgos, utilizando o algoritmo *CatBoostRegressor*. Essa conquista foi solidificada pela construção de uma robusta pipeline de preparação de dados, que envolveu a coleta e o tratamento de informações cruciais sobre as corridas e os galgos. Adicionalmente, o projeto incluiu a otimização sistemática dos hiperparâmetros do modelo *CatBoost*, resultando em uma configuração que maximiza a precisão das previsões. Finalmente, a pesquisa permitiu a definição de estratégias de aposta baseadas nas classificações projetadas dos galgos, demonstrando a aplicabilidade prática do modelo desenvolvido.

5.1 Conclusão

- A análise das regras e padrões das corridas de galgos do Reino Unido foi concluída com sucesso, permitindo a definição das condições ótimas para a seleção de corridas, como as categorias ’A’ ou ’OR’ e distâncias entre 460 e 500 metros, que representaram 50,1% das observações originais.
- O processo de definição e aplicação de uma pipeline estruturada de preparo de dados foi efetivamente demonstrado, resultando em uma melhoria considerável na qualidade da base de dados. Isso se traduziu na remoção de 6,8% de valores ausentes em finalPosition, 11 atributos removidos por irrelevância ou redundância, e o tratamento de 2,5% de registros com valor zero em time. O volume final da base de dados atingiu 153.735 observações, com 0 células ausentes e 0 linhas duplicadas.
- As técnicas de otimização de hiperparâmetros para o algoritmo *CatBoostRegressor* foram demonstradas com sucesso, ao aprimorar significativamente o desempenho do modelo preditivo. A otimização resultou em um RMSE de 0.31059 e um MAE de 0.23769 no conjunto de validação.
- A análise dos resultados e a identificação de estratégias de aposta foram realizadas com êxito, aplicando o modelo em 1.131 corridas nunca vistas. Para a estratégia *Win*, foi obtida uma precisão de 42,11% em 76 corridas elegíveis (com margem de 0,15 segundos). Na estratégia *Top 3*, a precisão alcançou 75,21% em 351 apostas aptas (com margem de 0,12 segundos). Para *Lay Placed*, obteve-se 82,2% de acerto em 62 apostas aptas (com margem de 0,36 segundos). Contudo,

o cálculo dos ganhos potenciais foi inviabilizado pela indisponibilidade de cotações de retorno, devido a novas leis esportivas.

5.2 Limitações

Contudo, o estudo enfrentou uma limitação significativa na etapa de avaliação do potencial financeiro das estratégias de aposta. Devido a novas regulamentações nacionais, especificamente a Lei nº 14.790, de 29 de dezembro de 2023 (Brasil, 2023) que entrou em vigor junto a Portaria nº 125, de 31 de dezembro de 2024 (Ministério do Esporte, 2024), as corridas de galgos não foram incluídas nas modalidades esportivas autorizadas para apostas de quota fixa. Consequentemente, as cotações de retorno para esses eventos tornaram-se indisponíveis no mercado regulamentado, impossibilitando a estimativa de ganhos e a análise de retorno sobre o investimento.

5.3 Trabalhos Futuros

Para futuras pesquisas e aprimoramentos, sugerem-se os seguintes trabalhos:

- **Aprofundamento em Estratégias de Aposta:** Desenvolver estratégias de aposta mais elaboradas, que possam incorporar elementos como gestão de banca, diversificação de apostas e análise de odds em cenários alternativos (se aplicáveis e eticamente viáveis).
- **Expansão e Refinamento de Atributos do Galgo:** Investigar a inclusão de mais dados do galgo (ex: peso, treinadores, linhagem) para enriquecer o conjunto de características preditivas do modelo.
- **Tratamento e Parametrização de Dados Textuais (*Remarks*):** Implementar técnicas avançadas de processamento para extrair insights valiosos dos campos de comentários (remarks), transformando-os em atributos numéricos valiosos para o modelo.

REFERÊNCIAS

BHAGAT, M.; BAKARIYA, B. A comprehensive review of cross-validation techniques in machine learning. *IJSAT-International Journal on Science and Technology*, IJSAT, v. 16, n. 1, 2025. Citado 2 vezes nas páginas 21 e 22.

BOROWSKI, P.; CHLEBUS, M. et al. *learning in the prediction of flat horse racing results in Poland*. [S.l.]: University of Warsaw, Faculty of Economic Sciences, 2021. Citado na página 24.

Brasil. *Lei nº 14.790, de 29 de dezembro de 2023*. 2023. Diário Oficial da União. Acesso em: 01 Mai. 2025. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2023-2026/2023/Lei/L14790.htm>. Citado 2 vezes nas páginas 43 e 48.

BUDACH, L. et al. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*, 2022. Citado na página 21.

da Costa, I. B.; MARINHO, L. B.; PIRES, C. E. S. Forecasting football results and exploiting betting markets: The case of “both teams to score”. *International Journal of Forecasting*, v. 38, n. 3, p. 895–909, 2022. ISSN 0169-2070. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207021001084>. Citado 2 vezes nas páginas 18 e 19.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018. Citado 3 vezes nas páginas 15, 19 e 20.

FIALHO, G.; MANHãES, A.; TEIXEIRA, J. P. Predicting sports results with artificial intelligence – a proposal framework for soccer games. *Procedia Computer Science*, v. 164, p. 131–136, 2019. ISSN 1877-0509. CENTERIS 2019 - International Conference on ENTERprise Information Systems / ProjMAN 2019 - International Conference on Project MANagement / HCist 2019 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050919322033>. Citado na página 19.

FRYE, M.; SCHMITT, R. H. Structured data preparation pipeline for machine learning-applications in production. *17th IMEKO TC*, v. 10, p. 241–246, 2020. Citado 3 vezes nas páginas 15, 20 e 21.

Gazeta do Povo. *Situação atual do mercado de apostas esportivas no Brasil*. 2023. Acesso em: 18 dez. 2024. Disponível em: <https://www.gazetadopovo.com.br/conteudo-publicitario/lottoland/situacao-atual-do-mercado-de-apostas-esportivas-no-brasil>. Citado na página 14.

Greyhound Board of Great Britain. *Greyhound Racing Data*. 2024. Acessado em: 31 de outubro de 2024. Disponível em: <https://www.gbgb.org.uk>. Citado na página 28.

GU, W. et al. A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, v. 130, p. 293–305, 2019. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417419302556>. Citado na página 19.

HANCOCK, J. T.; KHOSHGOFTAAR, T. M. Catboost for big data: an interdisciplinary review. *Journal of big data*, Springer, v. 7, n. 1, p. 94, 2020. Citado na página 19.

HUGGINS, M. “everybody’s going to the dogs”? the middle classes and greyhound racing in britain between the wars. *Journal of Sport History*, University of Illinois Press for The North American Society for Sport History, v. 34, n. 1, p. 96–120, 2007. Citado na página 17.

JAIN, A. et al. Overview and importance of data quality for machine learning tasks. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. [S.l.: s.n.], 2020. p. 3561–3562. Citado na página 20.

LAYBOURN, K. Going to the dogs: A history of greyhound racing in britain, 1926–2017. In: *Going to the dogs*. [S.l.]: Manchester University Press, 2019. Citado na página 14.

LYONS, A. *Man v Machine: Greyhound Racing Predictions*. Tese (Doutorado) — Dublin, National College of Ireland, 2016. Citado 4 vezes nas páginas 14, 15, 24 e 25.

Ministério do Esporte. *Portaria nº 125, de 31 de dezembro de 2024*. 2024. Diário Oficial da União. Acesso em: 01 Mai. 2025. Disponível em: <https://www.in.gov.br/web/dou/-/portaria-mesp-n-125-de-30-de-dezembro-de-2024-605034388>. Citado 2 vezes nas páginas 44 e 48.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, Volume 7 - 2013, 2013. Disponível em: <https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021>. Citado na página 19.

NTI, I. K. et al. Performance of machine learning algorithms with different k values in k-fold cross-validation. *International Journal of Information Technology and Computer Science*, MECS Publisher, v. 13, n. 6, p. 61–71, 2021. Citado 2 vezes nas páginas 21 e 22.

Oxford Stadium. *How Do Greyhound Racing Grades Work?* 2024. Accessed: 2024-11-10. Disponível em: <https://oxford-stadium.co.uk/blog/how-do-greyhound-racing-grades-work/>. Citado 2 vezes nas páginas 17 e 18.

Oxford Stadium. *What Information Is Included in a Greyhound Racecard?* 2024. Accessed: 2025-06-10. Disponível em: <https://oxford-stadium.co.uk/blog/what-information-is-included-in-a-greyhound-racecard/>. Citado na página 18.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. In: BENGIO, S. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. v. 31. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf>. Citado 3 vezes nas páginas 14, 19 e 20.

ROUDAUT, E. “tote clubs”, dog tracks and irish sweepstake: Controversy and compromise over popular gambling in interwar britain. *Angles. New Perspectives on the Anglophone World*, Société des Anglicistes de l’Enseignement Supérieur, n. 5, 2017. Citado na página 17.

SCHUMAKER, R. P. Machine learning the harness track: Crowdsourcing and varying race history. *Decision Support Systems*, v. 54, n. 3, p. 1370–1379, 2013. ISSN 0167-9236. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016792361200379X>. Citado na página 18.

SCHUMAKER, R. P.; JOHNSON, J. W. An investigation of svm regression to predict longshot greyhound races. *Communications of the IIMA*, v. 8, n. 2, p. 7, 2008. Citado 3 vezes nas páginas 15, 23 e 25.

SPORTS, C. *Aprendizado de Máquina na Análise de Esportes*. 2024. Acessado em: 7 dez. 2024. Disponível em: <https://www.catapult.com/pt/blogue/aprendizado-de-maquina-de-analise-de-esportes>. Citado na página 14.

Vaughan Williams, L.; STEKLER, H. O. Sports forecasting. *International Journal of Forecasting*, v. 26, n. 3, p. 445–447, 2010. ISSN 0169-2070. Sports Forecasting. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207009002064>. Citado na página 18.

WOLFF, M. B. *Garbage In, Garbage Out: The Importance of Good Data*. 2019. <https://medium.com/@marybrwolff/garbage-in-garbage-out-the-importance-of-good-data-ce1bb775468e>. Accessed: 2024-11-21. Citado na página 15.

WONG, T.-T.; YEH, P.-Y. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, v. 32, n. 8, p. 1586–1594, 2020. Citado na página 22.

ZHU, F. et al. Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*, 2024. Citado na página 18.

APÊNDICE A – FRAGMENTOS DO CÓDIGO-FONTE

Listing A.1 – Função de Treinamento do Modelo

```
def train_model_gridsearch(loss_function, scoring):
    feature_fields = ['timeMed', 'finMed', 'posLargMed', 'splitFinMed', 'recMed', 'distance', 'velMed', 'varMed', 'modTrap', 'trap', 'raceCat', 'ultCat', 'overall']
    target_field = 'time'

    # 1. Consulta no banco de dados
    qs = FinalDatabase.objects.all().values(*feature_fields, target_field)

    # Conversão para DataFrame
    df = pd.DataFrame(list(qs))

    # Separação X e y
    X = df[feature_fields]
    y = df[target_field]

    # 2. Divisão treino/validação
    X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

    # 3. Definição do modelo base
    cat_estimator = CatBoostRegressor(task_type='GPU', devices='0',
        random_seed=42, thread_count=6,
        loss_function=loss_function,
        eval_metric=loss_function,
        verbose=0)

    # 4. Configuração de K-Folds (4 folds)
    kf = KFold(n_splits=4, shuffle=True, random_state=42)

    # 5. Definição do grid de hiperparâmetros
    grid = {
        'iterations': [200, 500, 1000],      # número de árvores
        'learning_rate': [0.01, 0.03, 0.1],  # taxa de aprendizado
        'depth': [4, 6, 8],                  # profundidade das árvores
        'l2_leaf_reg': [2, 5, 9, 11],        # regularização L2
        'border_count': [32, 64, 128]        # número de bins para variáveis numéricas
    }
```

```
36
37      # 6. Configuração do GridSearchCV do Scikit-learn
38      grid_search_cv = GridSearchCV(estimator=cat_estimator, param_grid=grid,
         cv=kf, scoring=scoring, random_state=42, n_jobs=2, verbose=3, error_
        score='raise', return_train_score=True)
39
40      categorical_features = ['modTrap', 'trap', 'raceCat', 'ultCat']
41
42      # 7. Execução da busca de hiperparâmetros
43      fit_params_for_catboost = { 'cat_features': categorical_features,'early
        _stopping_rounds': 100}
44      grid_search_cv.fit(X_train, y_train, **fit_params_for_catboost)
45
46      # 8. Resultados da busca
47      best_params = grid_search_cv.best_params_
48      best_score_cv = -grid_search_cv.best_score_
49      print("Melhores parâmetros:", best_params)
50      print("Melhor RMSE / MAE médio: {:.4f}".format(best_score_cv))
51
52      # 9. Avaliação no conjunto de validação
53      best_model = grid_search_cv.best_estimator_
54      y_pred_val = best_model.predict(X_val)
55      rmse_val = np.sqrt(mean_squared_error(y_val, y_pred_val))
56      print("RMSE no conjunto de validação: {:.4f}".format(rmse_val))
57
58      # 10. (Opcional) Salvar o modelo otimizado
59      best_model.save_model(f'catboost_best_model_{loss_function}.cbm')
60      print(f"Modelo salvo como 'catboost_best_model_{loss_function}.cbm'")
```