



UFOP

Universidade Federal
de Ouro Preto

**Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas**

Aprendizado Residual Híbrido para Predição de Tempos de Volta e Insights Estratégicos em Múltiplos Circuitos de Fórmula 1

Marcos Paulo de Oliveira Pereira

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:

Alexandre Magno de Sousa

COORIENTAÇÃO:

Carlos Henrique Gomes Ferreira

Julho, 2026

João Monlevade—MG

Marcos Paulo de Oliveira Pereira

**Aprendizado Residual Híbrido para Predição de
Tempos de Volta e Insights Estratégicos em
Múltiplos Circuitos de Fórmula 1**

Orientador: Alexandre Magno de Sousa

Coorientador: Carlos Henrique Gomes Ferreira

Monografia apresentada ao curso de Engenharia da computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina “Trabalho de Conclusão de Curso II”.

Universidade Federal de Ouro Preto

João Monlevade

Julho de 2026

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

P436a Pereira, Marcos Paulo de Oliveira.

Aprendizado residual híbrido para predição de tempos de volta e insights estratégicos em múltiplos circuitos de Fórmula 1. [manuscrito] / Marcos Paulo de Oliveira Pereira. - 2026.

91 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Alexandre Magno Sousa.

Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Computação .

1. Aprendizado do computador. 2. Automóveis de Fórmula 1 - Corridas. 3. Corridas - Previsão. 4. Sistemas híbridos. I. Sousa, Alexandre Magno. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.85:796.7

Bibliotecário(a) Responsável: Flavia Reis - CRB6/2431



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTAÇÃO E SISTEMAS



FOLHA DE APROVAÇÃO

Marcos Paulo de Oliveira Pereira

Aprendizado Residual Híbrido para Predição de Tempos de Volta e Insights Estratégicos em Múltiplos Circuitos de Fórmula 1

Monografia apresentada ao Curso de Engenharia de Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação

Aprovada em 31 de julho de 2026

Membros da banca

Prof. Dr. Alexandre Magno de Sousa - Orientador (UFOP)
Prof. Dr. Carlos Henrique Gomes Ferreira - Coorientador (UFOP)
Profª. Drª. Helen de Cássia Sousa da Costa Lima (UFOP)
Prof. Dr. Luiz Carlos Bambirra Torres (UFOP)

Professor Alexandre Magno de Sousa, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/08/2026.



Documento assinado eletronicamente por **Alexandre Magno de Sousa, PROFESSOR DE MAGISTERIO SUPERIOR**, em 19/08/2026, às 14:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1168851** e o código CRC **EBFC8DED**.

Este trabalho é dedicado à todos aqueles que me apoiaram e me ajudaram a passar essa etapa importante da minha vida.

Agradecimentos

Agradeço primeiramente à minha família, pela base e direcionamento ao qual foi me dado desde meus primeiros passos. Agradeço à minha mãe, Suely, por ser minha porta de entrada no meio acadêmico e por nunca deixar de acreditar no meu potencial. Ao meu pai e exemplo, Sebastião, por todo seu carinho e amor pela família; e pelo esforço de suas mãos que me fizeram ter todas as oportunidades que eu poderia pedir. Aos meus irmãos, Pedro e João, que constituem uma parte importante de mim e me orientam com seus conselhos. À Alice, por todo seu suporte, atenção e companheirismo que me tornam mais forte e completo à cada dia, além de não me deixar desistir. Aos meus tios, Maria Vitalina e Ademir, por todo seu suporte e amor, e à minha madrinha, Alessandra, por seu carinho e cuidado especiais e aos seus incetivos à leitura desde cedo.

Agradeço aos meus companheiros e irmãos de casa, Lucas, Matheus, Nicolas e Otávio; o caminho ao qual trilhamos me construiu como pessoa e o apoio de vocês é especial. À Mariles, minha amiga e companheira de curso, que sempre me motiva a ser mais dedicado à cada dia. Ao Arthur, meu amigo e companheiro de curso, e à todos os outros companheiros que fizeram parte dessa etapa.

Agradeço ao meu orientador, Alexandre, principalmente por acreditar em mim e pelo seu esforço e ensinamentos durante a construção desta monografia. Ao Carlos, meu co-orientador, por também acreditar no trabalho e por seus ensinamentos em sala de aula. Agradeço também à todos os professores que, academicamente e pessoalmente, colaboraram na minha formação.

Por fim, deixo a minha eterna gratidão e o meu muito obrigado do fundo do coração.

"On a given day, a given circumstance, you think you have a limit. And you then go for this limit and you touch this limit, and you think, 'Okay, this is the limit.' As soon as you touch this limit, something happens, and you suddenly can go a little bit further. With your mind power, your determination, your instinct, and the experience as well, you can fly very high. Don't think you have limits."

— Ayrton Senna (1960 – 1994),
in: Senna (2010).

Resumo

A predição do tempo de volta é um componente central da análise estratégica na Fórmula 1, uma vez que pequenas variações no ritmo de corrida podem influenciar decisões como o momento dos *pit stops*, a escolha do composto de pneu e a viabilidade de táticas como o *undercut* e o *overcut*. Este trabalho propõe um *pipeline* de aprendizado de máquina para a predição de tempos de volta em múltiplos circuitos, utilizando dados públicos de telemetria extraídos por meio da biblioteca FastF1, cobrindo cinco circuitos (Bahrein, Arábia Saudita, Estados Unidos, Itália e Hungria) ao longo das temporadas de 2022 a 2025, período em que vigorou um único regulamento técnico. Para lidar com a limitação de dados por piloto e por pista, é proposta uma abordagem híbrida de dois estágios, denominada MLR-LSTM Híbrido, na qual um modelo de Regressão Linear Múltipla estima a tendência agregada do ritmo de corrida ao longo das temporadas, enquanto uma rede Long Short-Term Memory modela os resíduos desse modelo, filtrados por piloto e por ano, capturando dependências temporais não lineares. O modelo híbrido é comparado a modelos de Regressão Linear e XGBoost, sendo avaliado por meio de uma validação temporal baseada em janela expansiva e de um indicador de *score* composto COS, que pondera desempenho médio e variância entre treino e teste. Diferentemente de abordagens que tratam a predição como fim em si mesma, todos os três modelos são submetidos a uma análise interpretativa, apoiada em coeficientes de regressão, importância de *features* e valores SHAP. Essa análise permite identificar quais fatores climáticos, mecânicos e temporais mais influenciam o ritmo de corrida em cada circuito, e comparar como cada modelo (linear, não linear e híbrido) captura essas relações de forma distinta. Os resultados mostram que a abordagem híbrida melhora a predição em cenários de dados limitados, notadamente na Itália, e que a extensão da interpretabilidade também ao componente residual do modelo híbrido permite extrair insights estratégicos específicos para cada Grande Prêmio, oferecendo uma base tanto para a predição do ritmo de corrida quanto para a compreensão de informações importantes quanto às estratégias.

Palavras-chaves: Predição de Tempo de Volta. Fórmula 1. Modelo Híbrido. LSTM. Insights Estratégicos. Interpretabilidade.

Abstract

Lap time prediction is a central component of strategic analysis in Formula 1, since small variations in race pace can influence decisions such as pit stop timing, tire compound selection, and the viability of tactics such as the undercut and the overcut. This work proposes a machine learning pipeline for lap time prediction across multiple circuits, using public telemetry data extracted through the FastF1 library, covering five circuits (Bahrain, Saudi Arabia, United States, Italy, and Hungary) across the 2022 to 2025 seasons, a period in which a single technical regulation was in effect. To address data limitations per driver and per track, a two-stage hybrid approach is proposed, referred to as Hybrid MLR-LSTM, in which a Multiple Linear Regression model estimates the aggregated race pace trend across seasons, while a Long Short-Term Memory network models this model's residuals, filtered by driver and by year, capturing nonlinear temporal dependencies. The hybrid model is compared against Linear Regression and XGBoost models, being evaluated through a temporal validation based on expanding window and a composite score indicator COS, which weights average performance and variance between training and testing. Unlike approaches that treat prediction as an end in itself, all three models are subjected to a common interpretive analysis, supported by regression coefficients, feature importance and SHAP values. This analysis allows identifying which climatic, mechanical, and temporal factors most influence race pace on each circuit, and comparing how each model (linear, nonlinear, and hybrid) captures these relationships differently. The results show that the hybrid approach improves prediction in data-limited scenarios, notably in Italy, and that extending interpretability to the hybrid model's residual component also allows the extraction of circuit-specific strategic insights, providing a basis for both race pace prediction and the understanding of the key factors regarding strategies.

Keywords: Lap Time Prediction. Formula 1. Hybrid Model. LSTM. Strategic Insights. Interpretability.

Lista de ilustrações

Figura 1 – Exemplificação da escolha de compostos pirelli.	25
Figura 2 – Stints durante o Grande Prêmio do Bahrein em 2025.	25
Figura 3 – Representação do comportamento de um composto de pneu arbitrário durante uma corrida de Fórmula 1.	26
Figura 4 – Representação da aerodinâmica de um carro de F1.	26
Figura 5 – Traçado do Circuito Internacional do Bahrein.	27
Figura 6 – Traçado do Circuito Corniche de Jeddah.	28
Figura 7 – Traçado do Circuito das Américas.	28
Figura 8 – Traçado do Circuito Nacional de Monza.	29
Figura 9 – Traçado do Hungaroring.	30
Figura 10 – Pipeline metodológico experimental.	37
Figura 11 – Matriz de correlação.	40
Figura 12 – Loadings PCA.	41
Figura 13 – Distribuição das <i>features</i> de condições climáticas e do tempo de volta.	43
Figura 14 – Número de voltas por composto de pneu e ano em cada GP.	44
Figura 15 – Detalhes das configurações dos experimentos <i>baselines</i>	45
Figura 16 – Abordagem Híbrida de Dois Estágios.	47
Figura 17 – CDF das métricas por piloto para o Grande Prêmio da Arábia Saudita.	55
Figura 18 – CDF das métricas por piloto para o Grande Prêmio dos Estados Unidos.	55
Figura 19 – CDF das métricas por piloto para o Grande Prêmio da Itália.	55
Figura 20 – Desempenho agregado no conjunto de teste (<i>holdout</i> de 2025) por circuito e modelo.	56
Figura 21 – Séries temporais comparativas para o <i>holdout</i> na Hungria para o piloto Norris.	57
Figura 22 – Séries temporais comparativas para o <i>holdout</i> na Itália para o piloto Verstappen.	58
Figura 23 – Séries temporais comparativas para o <i>holdout</i> os EUA para o piloto Russell.	59
Figura 24 – Densidade do tempo de volta em cada um dos circuitos.	59
Figura 25 – Coeficientes da Regressão Linear Múltipla.	60
Figura 26 – Ganho de importância no XGBoost.	61
Figura 27 – SHAP <i>Values</i> para o Grande Prêmio da Arábia Saudita.	62
Figura 28 – Decomposição SHAP (<i>force plot</i>) da predição de uma volta no GP da Arábia Saudita no <i>holdout</i>	62
Figura 29 – SHAP <i>Values</i> para o Grande Prêmio da Itália no modelo LSTM.	63

Figura 30 – Decomposição SHAP (<i>force plot</i>) da predição de uma volta no GP da Itália no <i>holdout</i>	64
Figura 31 – Perfil de recência para o modelo LSTM.	64

Lista de tabelas

Tabela 1 – Resumo dos trabalhos relacionados, organizados segundo as quatro vertentes discutidas na fundamentação teórica.	33
Tabela 2 – Dicionário de dados do dataset.	38
Tabela 3 – Quantidade de voltas registradas por pista, antes e após a filtragem dos dados.	38
Tabela 4 – Espaço de busca do Optuna para o XGBoost-EW por pista.	46
Tabela 5 – Hiperparâmetros selecionados para o XGBoost-EW (mediana entre folds) por pista.	46
Tabela 6 – Hiperparâmetros selecionados para o LSTM, por pista.	48
Tabela 7 – Comparação do desempenho dos modelos filtrados por piloto e LSTM puro.	53
Tabela 8 – Comparação do desempenho entre os modelos.	54
Tabela 9 – Síntese dos <i>insights</i> estratégicos por circuito.	66
Tabela 10 – Tabela de sumarização estatística dos dados para a pista de Bahrein.	77
Tabela 11 – Tabela de sumarização estatística dos dados para a pista de Arábia Saudita.	77
Tabela 12 – Tabela de sumarização estatística dos dados para a pista de Hungria.	77
Tabela 13 – Tabela de sumarização estatística dos dados para a pista de Itália.	78
Tabela 14 – Tabela de sumarização estatística dos dados para a pista de Estados Unidos.	78
Tabela 15 – Tabela de sumarização estatística dos dados brutos para a pista de Bahrein (voltas: 4438 registros; clima: 639 registros).	78
Tabela 16 – Tabela de sumarização estatística dos dados brutos para a pista de Arábia Saudita (voltas: 3562 registros; clima: 588 registros).	78
Tabela 17 – Tabela de sumarização estatística dos dados brutos para a pista de Hungria (voltas: 5358 registros; clima: 643 registros).	79
Tabela 18 – Tabela de sumarização estatística dos dados brutos para a pista de Itália (voltas: 3912 registros; clima: 580 registros).	79
Tabela 19 – Tabela de sumarização estatística dos dados brutos para a pista de Estados Unidos (voltas: 4132 registros; clima: 647 registros).	79
Tabela 20 – Média do tempo de volta (s) por piloto e por ano – Bahrein.	80
Tabela 21 – Média do tempo de volta (s) por equipe e por ano – Bahrein.	80
Tabela 22 – Média do tempo de volta (s) por piloto e por ano – Arabia Saudita.	81
Tabela 23 – Média do tempo de volta (s) por equipe e por ano – Arabia Saudita.	81
Tabela 24 – Média do tempo de volta (s) por piloto e por ano – Hungria.	82
Tabela 25 – Média do tempo de volta (s) por equipe e por ano – Hungria.	82

Tabela 26 – Média do tempo de volta (s) por piloto e por ano – Italia.	83
Tabela 27 – Média do tempo de volta (s) por equipe e por ano – Italia.	83
Tabela 28 – Média do tempo de volta (s) por piloto e por ano – Estados Unidos. . .	84
Tabela 29 – Média do tempo de volta (s) por equipe e por ano – Estados Unidos. .	84
Tabela 30 – Posição de chegada de cada piloto por ano – Arabia Saudita.	85
Tabela 31 – Posição de chegada de cada piloto por ano – Hungria.	86
Tabela 32 – Posição de chegada de cada piloto por ano – Italia.	86
Tabela 33 – Posição de chegada de cada piloto por ano – Estados Unidos.	87
Tabela 34 – Coeficientes da regressão linear (janela expansiva). Piloto, equipe e composto omitidas para melhor exibição.	88
Tabela 35 – Importância das variáveis no XGBoost (janela expansiva). Piloto, equipe e composto omitidas para melhor exibição.	88
Tabela 36 – Importância média absoluta dos valores SHAP do XGBoost no conjunto de teste (<i>holdout</i> sequencial), por circuito.	89
Tabela 37 – Importância média absoluta dos valores SHAP do modelo híbrido (componente LSTM) no conjunto de teste (<i>holdout</i> sequencial), por circuito.	89
Tabela 38 – Desvio padrão dos resíduos (s) no treino (validação) e no teste (holdout sequencial) por circuito – Regressao Linear (janela expansiva).	89
Tabela 39 – Desvio padrão dos resíduos (s) no treino (validação) e no teste (holdout sequencial) por circuito – XGBoost (janela expansiva).	89
Tabela 40 – Desvio padrão dos resíduos (s) no treino (validação) e no teste (holdout sequencial) por circuito – Modelo Híbrido.	90

Lista de abreviaturas e siglas

FIA	Federação Internacional do Automóvel
F1	Fórmula 1
LSTM	Long Short-Term Memory
MLR	Multiple Linear Regression
XGBoost	Extreme Gradient Boosting
COS	Composite Overfitting Score
PCA	Principal Component Analysis
SHAP	Shapley Additive Explanations
EW	Expanding Window
SW	Sliding Window
QP	Questão de Pesquisa
GP	Grande Prêmio
ML	Machine Learning
RBF	Gaussian Radial Basis Function

Sumário

1	INTRODUÇÃO	16
1.1	Motivação e Justificativa	17
1.2	Objetivo Geral e Questões de Pesquisa	18
1.3	Resultados e Contribuições	19
1.4	Estrutura e Organização da Monografia	21
2	FUNDAMENTAÇÃO TEÓRICA	22
2.1	Fundamentos da Temporada de Fórmula 1	22
2.2	Fundamentos do Ritmo de Corrida na Fórmula 1	23
2.3	Circuitos Analisados	27
2.3.1	Circuito Internacional do Bahrein	27
2.3.2	Circuito Corniche de Jeddah	27
2.3.3	Circuito das Américas	28
2.3.4	Circuito Nacional de Monza	29
2.3.5	Hungaroring	29
2.4	Trabalhos Relacionados	30
2.5	Considerações Finais	34
3	METODOLOGIA	36
3.1	Preparação dos Dados	36
3.1.1	Extração de Dados	36
3.1.2	Pré-processamento de Dados	37
3.1.3	Engenharia de Features e Análise Exploratória de Dados	39
3.2	Configuração dos Experimentos	45
3.3	Algoritmos Baseline e Abordagem de Aprendizado Profundo em Dois Estágios	45
3.4	Abordagem de Score Balanceado para Seleção de Modelos	49
3.5	Interpretabilidade dos Modelos	49
3.6	Considerações Finais	50
4	RESULTADOS	52
4.1	Comparação de Desempenho entre Modelos	52
4.2	Análise Temporal do Holdout	56
4.3	Análise Interpretativa dos Modelos	59
4.4	Insights Estratégicos por Circuito	64
4.5	Considerações Finais	65

5	CONCLUSÃO	68
5.1	Síntese da Metodologia Utilizada	68
5.2	Discussão sobre as Questões de Pesquisa	68
5.3	Limitações do Trabalho	70
5.4	Principais Contribuições	70
5.5	Trabalhos Futuros	71
	REFERÊNCIAS	73
	APÊNDICES	76
	APÊNDICE A – ANÁLISE EXPLORATÓRIA DA CARACTERÍSTICA DAS PISTAS	77
A.1	Sumarização Estatística para os Dados Filtrados	77
A.2	Sumarização Estatística para os Dados Brutos	78
A.3	Tempo de Volta ao Longo dos Anos	80
	APÊNDICE B – RESULTADOS DAS CORRIDAS EM CADA ANO DO DATASET	85
B.1	Resultados das Corridas	85
	APÊNDICE C – ANÁLISE DOS RESULTADOS DOS MODELOS	88
C.1	Coeficientes da Regressão Linear	88
C.2	Importância das Variáveis	88
C.3	SHAP <i>Values</i>	89
C.4	Análise de Resíduos	89

1 INTRODUÇÃO

A categoria de corrida de Fórmula 1 (F1) é reconhecida de forma inquestionável como o auge do automobilismo mundial ([KRZYSZTOŃ; SMOLKA, 2026](#); [STEPIEN, 2025](#)). Reúne vinte dos melhores pilotos do mundo dirigindo carros de tecnologia de ponta em corridas ao longo de mais de 20 países e 4 continentes. Com mais de 70 anos de história, a F1 atravessou eras marcadas por profundas transformações técnicas e regulamentares: dos motores aspirados às unidades de potência híbridas, dos chassis de alumínio aos monocoques de fibra de carbono, e de um esporte guiado quase que exclusivamente pela intuição de pilotos e engenheiros para uma categoria em que cada decisão é sustentada por dados. A cada novo ciclo de regulamento técnico, as equipes são forçadas a reinventar seus projetos, o que faz da categoria um ambiente de evolução contínua. No entanto, essa categoria não é apenas esporte e entretenimento, ela tem uma importância significativa para a sociedade, pois sua tecnologia tem grandes impactos em veículos de passeio, tais como evoluções mecânicas, desenvolvimento de combustíveis sustentáveis, motores híbridos, inovações aerodinâmicas e segurança ([Fédération Internationale de l'Automobile, 2025](#); [STEPIEN, 2025](#)). A F1 é muitas vezes referenciada como um grande laboratório das montadoras de carros mais famosas do mundo, como a Mercedes, a McLaren, a Ford (através da Red Bull Racing) e a Ferrari. Tecnologias hoje comuns em carros de rua, como sistemas de recuperação de energia, materiais leves de alta resistência e avanços em segurança estrutural, tiveram nessa categoria de corrida um de seus principais campos de prova. Esse ciclo de transferência tecnológica reforça que os investimentos e as inovações da F1 extrapolam o ambiente das pistas, alcançando a indústria automotiva como um todo.

Além das informações anteriormente mencionadas, a F1 também se destaca pelo volume massivo de dados gerados e analisados em tempo real. Pode-se dizer que a categoria é orientada por testes, cada volta em um treino, cada simulação no computador. Todas essas ocasiões geram dados e dados geram mudanças ([BANSAL et al., 2025](#); [THOMAS et al., 2025](#)). Um carro moderno de F1 é equipado com centenas de sensores que monitoram continuamente parâmetros como temperatura dos pneus e dos freios, pressões, cargas aerodinâmicas e comportamento da suspensão, transmitindo essas informações em tempo real para os engenheiros no muro dos boxes e nas fábricas. Decisões estratégicas são tomadas pela equipe técnica que recebe todo esse volume de dados capturados pelo sistema de telemetria veicular para que possam solucionar problemas críticos do dia-a-dia do automobilismo da F1 ([FATIMA; JOHRENDT, 2023](#); [GARCÍA, 2024](#)).

Dentre esses problemas, podem ser citados aqueles relacionados aos pneus, como

o *pit stop*, o momento em que o piloto pára nos boxes para que a equipe de mecânicos efetue a troca de pneus do carro, e a gestão de pneus, que diz respeito a como contornar a durabilidade dos diferentes tipos de pneu durante uma corrida, escolhendo o mais veloz e com durabilidade adequada a cada momento. A escolha do momento da parada é particularmente sensível: parar cedo demais expõe o piloto ao tráfego e a um longo *stint* final, enquanto parar tarde demais pode significar perder posição para um rival que se beneficiou de pneus novos. Também é possível mencionar dados que auxiliam diretamente as equipes no desenvolvimento do carro, como informações do comportamento em curvas de alta, média e baixa velocidade, e a potência e velocidade máxima do carro em retas. Todos esses dados alteram as direções de desenvolvimento aerodinâmico e do desenvolvimento do motor. Além disso, através dos treinos, classificações, corridas e até mesmo simuladores nas sedes das equipes, os próprios pilotos podem adquirir *feedback* direto do seu desempenho (HEILMEIER et al., 2020; HEILMEIER; GRAF; LIENKAMP, 2018; THOMAS et al., 2025).

1.1 Motivação e Justificativa

Para a corrida, a previsão do ritmo de corrida, do inglês, *race pace*, é vista como um dos fatores mais importantes que ditam a impressão do final de semana que envolve o treino livre, a corrida *sprint* e a corrida principal (GARCÍA, 2024). O ritmo de corrida é um problema complexo que é influenciado por variáveis não lineares e estocásticas, envolvendo a degradação dos pneus por composto, as condições climáticas e as diferenças sutis entre a dirigibilidade, seja de um piloto ou de um carro específico (HEILMEIER et al., 2020; HEILMEIER; GRAF; LIENKAMP, 2018; FATIMA; JOHRENDT, 2023). Diante desse cenário, é evidente que esse problema necessita de bastante cuidado para ser modelado de forma que responda a questões referentes ao desempenho dos pilotos, à evolução do veículo e às nuances dos circuitos.

Para obter estratégias eficazes e robustas, é essencial considerar os efeitos de eventos aleatórios, como, por exemplo, *safety cars*, que afetam o curso da corrida (HEILMEIER et al., 2020). No entanto, essa aleatoriedade é difícil de ser modelada, o que torna a elaboração desse tipo de modelo um pouco mais cuidadosa. Nesse contexto, técnicas de aprendizado de máquina surgem como uma ferramenta poderosa, capaz de identificar padrões complexos em históricos de dados e gerar modelos preditivos (CHETELES, 2024; TODD et al., 2025). Um modelo que consiga prever o tempo de volta ideal poderia ser integrado em uma simulação incorporada com fatores de efeito aleatório, funcionando como um agente que auxiliaria as decisões em tempo real, como pode ser visto no contexto da Fórmula-E no trabalho de (LIU; FOTOUHI, 2020).

Além disso, como as condições climáticas variam significativamente entre os países

e cada circuito da temporada da F1 tem tamanho variável, para prever os tempos de volta de um piloto, é essencial separar a série temporal de cada piloto por circuito (ZHAO, 2024). Essa exigência reduz a quantidade de dados disponíveis e pode tornar o modelo menos preciso (KARA; KüçüKMANISA, 2025).

1.2 Objetivo Geral e Questões de Pesquisa

O objetivo geral deste trabalho é desenvolver e validar um modelo de predição capaz de prever tempos de volta com precisão, contornando cenários de limitação de dados através da elaboração de um *pipeline* de ciência de dados. Ainda, o modelo selecionado para cada contexto precisa considerar desempenho e variância. Além disso, este estudo tem como objetivo extrair informações suficientes da saída dos modelos para que se trace um perfil estratégico de cada pista estudada, entendendo os fatores climáticos, físicos e temporais de cada circuito que influenciam o tempo de volta.

Nesse contexto, o problema exige uma abordagem precisa para responder às seguintes Questões de Pesquisa (QP):

- **QP1:** *Como o tempo de volta de um piloto de F1 pode ser predito em cenários com dados limitados?*

Como separar as séries temporais de tempos de volta por circuito para cada piloto reduz e limita os dados disponíveis, uma solução possível é agregar os tempos de volta de todos os pilotos em um determinado circuito em uma única série temporal combinada, abrangendo múltiplas temporadas;

- **QP2:** *Quais estratégias podem melhorar a acurácia dos modelos de predição do tempo de volta de pilotos?*

Esse trabalho propõe um modelo híbrido para prever tempos de voltas. Essa abordagem utiliza um modelo de aprendizado profundo para corrigir as predições de um modelo baseline. Mais especificamente, uma rede neural profunda Long Short-Term Memory (LSTM) é utilizada para prever os resíduos do modelo baseline, que tem como alvo o tempo de volta. A previsão do modelo híbrido é então obtida ao somar a predição do baseline com os resíduos preditos pelo LSTM;

- **QP3:** *Como o modelo mais adequado pode ser selecionado equilibrando desempenho e variância do modelo?*

Diversas métricas de desempenho conseguem avaliar modelos de aprendizado de máquina, no entanto, é importante equilibrar as médias de desempenho com a variância do modelo. Ao final, este trabalho propõe uma métrica indicativa que combina esses dois fatores;

- **QP4:** *O que as informações de explicabilidade provenientes de cada modelo revelam sobre as características específicas para a estratégia de cada circuito, considerando atributos climáticos, mecânicos e temporais?*

Com o objetivo de compreender em que medida cada circuito difere dos demais e validar a abordagem multi-circuito, este trabalho busca através de técnicas de explicabilidade compreender e quantificar o impacto de cada variável nos circuitos estudados para cada modelo. É necessário cruzar as informações obtidas na análise exploratória com resultados diretos obtidos da explicabilidade dos modelos. Por meio desse conjunto de informações, o trabalho busca extrair *insights* estratégicos que podem ajudar a caracterizar a melhor estratégia para cada cenário de corrida.

A estrutura ditada pelas questões de pesquisa enfatiza que este trabalho não trata a predição do tempo de volta como um fim em si mesma. O problema do *race pace* vai além: o desafio de prever com precisão é retratado entre as questões de QP1 e QP3, mas a última questão, QP4, avança para além dessa predição para investigar o que ela revela sobre o comportamento de cada circuito. Dessa maneira, os *insights* estratégicos que podem ser obtidos através de cada resultado são tratados de maneira igualmente valiosa, construindo um objetivo de não apenas chegar ao resultado, mas também de poder explicá-lo.

1.3 Resultados e Contribuições

Os resultados finais confirmam a capacidade da abordagem híbrida proposta para lidar com a limitação de dados no problema de predição do ritmo de corrida. No conjunto de treino, o modelo Híbrido apresentou o melhor desempenho em todos os cinco circuitos analisados, superando tanto a Regressão Linear Múltipla, do inglês, Multiple Linear Regression (MLR), quanto o Extreme Gradient Boosting (XGBoost). Para o conjunto de teste, no circuito da Itália, o modelo híbrido apresentou desempenho significativamente superior aos demais em todas as métricas avaliadas, com ganhos de pelo menos 11% no RMSE e 27% no MAE em relação à Regressão Linear e ao XGBoost, respectivamente. Nos Estados Unidos, o modelo híbrido também apresentou o melhor equilíbrio entre desempenho médio e variância, segundo o indicador Composite Overfitting Score (COS) proposto neste trabalho. Já em circuitos como Bahrein e Hungria, observou-se um empate estatístico entre os modelos, o qual foi resolvido em favor da Regressão Linear Múltipla pelo critério de variância. Na Arábia Saudita, o XGBoost obteve a melhor relação entre desempenho e estabilidade. Esses resultados indicam que a superioridade do modelo híbrido não é uniforme entre circuitos, mas está associada às características específicas de cada traçado de pista e à forma como o ritmo de corrida se comporta ao longo das temporadas analisadas.

Além da predição, a análise interpretativa revelou padrões relevantes sobre os fatores que explicam o ritmo de corrida em cada circuito. O tempo de volta anterior se manteve como a variável mais consistente entre todos os circuitos, tanto na análise de coeficientes da Regressão Linear Múltipla quanto na importância de *features* e nos valores Shapley Additive Explanations (SHAP) do XGBoost, confirmando a forte dependência temporal do problema. Por outro lado, a relevância das demais variáveis variou conforme o circuito: (i) na Itália, o atributo referente ao ano da corrida se destacou como um dos principais fatores no XGBoost, associado à evolução dos carros entre temporadas em um circuito de baixa carga aerodinâmica; (ii) no Bahrein e nos Estados Unidos, a temperatura da pista assumiu papel de destaque, com efeitos opostos entre os dois circuitos; (iii) já na Arábia Saudita e na Hungria, fatores mais ligados à degradação, como o número da volta e a vida do pneu, se mostraram mais relevantes. Para o resíduo do LSTM, é interessante observar que seu resultado para a Itália não é coincidência: a análise dos SHAP *Values* da validação e do teste aponta que o modelo de aprendizado profundo dá importância para variáveis que a Regressão Linear despreza, como o composto de pneu, a vida do pneu e a velocidade do vento. Esses achados reforçam que cada Grande Prêmio (GP) deve ser tratado como um problema de modelagem específico, e que a extensão dessa análise interpretativa ao componente residual do modelo híbrido permite identificar nuances que os modelos puramente lineares ou baseados em árvores não capturam isoladamente.

A partir desses resultados, este trabalho apresenta as seguintes contribuições:

- (i) a proposta de um modelo híbrido de dois estágios com aprendizado residual (MLR-LSTM), capaz de mitigar a limitação de dados decorrente da separação da série temporal por piloto e por circuito;
- (ii) a proposta de um indicador de score composto (COS), que equilibra desempenho médio e variância entre treino e teste na seleção do modelo mais adequado para cada circuito;
- (iii) uma metodologia de interpretabilidade aplicada de forma unificada aos três tipos de modelo, estendendo a análise também ao componente residual aprendido pelo LSTM;
- (iv) uma análise multi-circuito que evidencia quais *features* mais impactam a saída do modelo, reforçando a necessidade de tratar cada Grande Prêmio como um contexto de modelagem próprio;
- (v) a sumarização da explicabilidade de cada modelo, resultando em *insights* estratégicos únicos referentes a fatores climáticos, mecânicos e temporais em cada um dos circuitos estudados.

1.4 Estrutura e Organização da Monografia

No Capítulo 2 são pontuadas informações a respeito dos principais conceitos envolvendo a temporada de F1 e a predição do ritmo de corrida, fundamento a base para decisões metodológicas futuras. Além disso, são apresentados os circuitos que serão analisados, com suas características individuais únicas. Em seguida, no Capítulo 3 é apresentada a metodologia que representa o cerne deste trabalho, esse capítulo é dividido em seis seções principais, o capítulo descreve desde a obtenção dos dados até a configuração dos modelos e a escolha das ferramentas de interpretabilidade.

Os resultados são apresentados no Capítulo 4, divididos em cinco seções. A primeira diz respeito a comparação geral do desempenho dos modelos, ajudando a responder as três primeiras questões de pesquisa. Em seguida, os resultados obtidos na primeira seção são expandidos focando na temporalidade do conjunto de teste, i.e. *holdout*. O capítulo fecha respondendo as últimas duas questões de pesquisa e faz uma análise interpretativa dos modelos bem como extrai os *insights* estratégicos de cada circuito.

Por fim, o Capítulo 5 trata das conclusões, limitações, contribuições e trabalhos futuros. Além disso, o trabalho conta com apêndices. O Apêndice A apresenta uma análise exploratória das características de cada um dos circuitos estudados, através de tabelas de sumarização estatística e caracterização do tempo de volta ao longo dos anos. O Apêndice B apresenta os resultados para cada corrida nos anos do *dataset*. Por fim, o apêndice C expande a análise dos resultados dos modelos, principalmente na parte da explicabilidade.

2 Fundamentação Teórica

Este capítulo apresenta uma revisão da literatura, bem como trabalhos correlatos, além de fundamentar a base necessária para entender o ritmo de corrida como um problema na Fórmula 1. A Seção 2.1 detalha o funcionamento de uma temporada de Fórmula 1, tais como a estrutura de um Grande Prêmio e o regulamento técnico da categoria. A Seção 2.2 se aprofunda nos conceitos da Fórmula 1 como categoria, principalmente no problema do *race pace*. Na Seção 2.3 são apresentados os circuitos utilizados para o estudo deste trabalho, junto das suas características e detalhes. Por sua vez, a Seção 2.4 é dedicada a apresentar os trabalhos relacionados existentes na literatura, classificando-os em diferentes vertentes de pesquisa. Por fim, a Seção 2.5 apresenta as considerações finais deste capítulo.

2.1 Fundamentos da Temporada de Fórmula 1

Antes de entender a vertente que o trabalho se propõe a explorar, é necessário compreender o funcionamento da F1. A Fórmula 1 é organizada em temporadas anuais, compostas por uma sequência de Grandes Prêmios disputados em circuitos ao redor do mundo. O calendário atual conta com mais de vinte etapas por ano, distribuídas entre Europa, Américas, Oriente Médio, Ásia e Oceania. Em cada temporada, dez equipes, também usualmente chamadas de construtores, inscrevem dois carros cada, totalizando vinte pilotos titulares no grid. Ao longo do ano, são disputados dois campeonatos em paralelo: o Campeonato Mundial de Pilotos, decidido pela pontuação individual, e o Campeonato Mundial de Construtores, decidido pela soma dos pontos dos dois carros de cada equipe. Os pontos são distribuídos entre os dez primeiros colocados de cada corrida. Com tanto investimento, tecnologia e qualidade envolvidas, cada décimo de segundo disputado em pista têm impacto direto no resultado esportivo e financeiro das equipes.

Cada Grande Prêmio é estruturado como um fim de semana de atividades. Tradicionalmente, nas sextas e sábados, acontecem as três sessões de treinos livres, do inglês Free Practice, nas quais as equipes coletam dados, testam configurações de carro e avaliam o comportamento dos pneus naquele circuito. No sábado, após o terceiro treino, ocorre uma sessão de classificação, na qual os pilotos buscam a volta mais rápida possível para definir a ordem de largada. É no treino classificatório que os carros de Fórmula 1 se encontram na sua janela de maior desempenho: pneus novos e baixa quantidade de combustível. Por fim, a corrida principal, realizada geralmente no domingo, distribui a pontuação do campeonato. Em alguns fins de semana selecionados, o formato inclui ainda uma corrida *sprint*, uma prova mais curta e com dinâmica própria. Como mencionado, essas sessões têm naturezas muito distintas entre si: uma volta de classificação representa o desempenho

máximo do conjunto carro-piloto, enquanto a corrida principal exige a sustentação de um ritmo competitivo por dezenas de voltas, sob gestão de pneus e combustível.

As regras que regem a categoria são definidas pela Federação Internacional do Automóvel (FIA) e se dividem em regulamento esportivo, que disciplina a condução do campeonato e das corridas, e regulamento técnico, que especifica as características construtivas dos carros. O regulamento técnico evolui em ciclos: periodicamente, a FIA introduz mudanças profundas de conceito, como a transição para as unidades de potência híbridas em 2014 ou a reformulação aerodinâmica baseada em efeito solo introduzida em 2022, que forçam as equipes a projetar carros substancialmente novos ([Fédération Internationale de l'Automobile, 2025](#)). É natural que dentro de um mesmo ciclo os carros evoluam de forma incremental. As equipes focam em atualizações de desenvolvimento, com pacotes com novos detalhes aerodinâmicos ou de potência se estendendo por toda a temporada. No entanto, é natural que dentro desse regulamento técnico se preserve uma base técnica comum. Essa dinâmica tem consequência direta para a análise de dados: comparar tempos de volta de regulamentos distintos significa comparar carros de gerações diferentes, com níveis de desempenho incompatíveis entre si. Portanto, é necessário um cuidado no momento de decidir quais dados utilizar e para o quê utilizar. Se o contexto é a predição direta de alguma elemento, como o tempo de volta ou estratégia ótima de parada, faz mais sentido que o *dataset* se concentre em um mesmo contexto de regulamento técnico.

Por fim, é bom ressaltar a escala de tempo normalmente utilizada na F1. Os cenários apresentados por uma temporada de Fórmula 1 sugerem um grid altamente competitivo, com tecnologia de ponta e os melhores pilotos do mundo. Dessa maneira, as diferenças de desempenho entre carros são frequentemente medidas em décimos ou centésimos de segundo por volta: uma vantagem consistente dessa magnitude, acumulada ao longo de dezenas de voltas, é suficiente para alterar posições de pista, janelas de parada e resultados de corrida. Essa escala ajuda a entender o padrão de exigência para modelos de predição de tempo de volta no contexto da categoria: erros médios que estejam na casa de poucos décimos de segundo situam-se na mesma ordem de grandeza das diferenças que decidem disputas reais.

2.2 Fundamentos do Ritmo de Corrida na Fórmula 1

O *race pace*, ou ritmo de corrida, é o tempo de volta que um piloto é capaz de sustentar de forma consistente ao longo de um estágio da corrida, em contraste com o tempo de uma volta isolada em condições de baixo combustível e pneus novos, como ocorre na qualificação, ou o tempo de volta lento durante uma corrida que esteja em condições adversas, como regime de *safety car* ou até mesmo sob chuva, se estamos nos referindo

ao ritmo puro em pista seca. Diferentemente da volta de classificação, o ritmo de corrida é resultado da interação entre múltiplos fatores que evoluem continuamente durante o Grande Prêmio: a degradação dos pneus, a redução da massa de combustível, as condições climáticas e da pista. Além disso, voltas com o carro de segurança na pista tem limites de velocidade por segurança; portanto, os pilotos são obrigados a reduzir deliberadamente seu próprio ritmo. No contexto da chuva, a velocidade real do carro cai drasticamente em razão das condições adversas da pista. Dessa maneira, quando o ritmo de corrida é mencionado neste trabalho, o ritmo puro é referido como o pacote carro e piloto consegue imprimir durante uma corrida, sendo esse próximo daquilo que se considera ideal. Compreender os fatores que impactam esse *race pace* é essencial para justificar as variáveis utilizadas na construção deste trabalho, uma vez que cada uma delas busca capturar um mecanismo físico ou estratégico específico que influencia o tempo de volta.

Um dos principais determinantes do ritmo de corrida é o comportamento dos pneus. A Pirelli, fornecedora exclusiva de pneus da Fórmula 1, disponibiliza um conjunto de compostos que variam de C1 (o mais duro) a C5 (o mais macio), sendo que, para cada fim de semana de corrida, três desses compostos são selecionados e nomeados relativamente como *hard*, *medium* e *soft*. Compostos mais macios oferecem maior aderência e, consequentemente, tempos de volta mais rápidos, mas degradam-se mais rapidamente, apresentando queda de desempenho ao longo do *stint* (o intervalo entre duas paradas nos boxes). Compostos mais duros, por sua vez, sacrificam desempenho de pico em favor de maior durabilidade. A Figura 1 ajuda a entender visualmente como funciona a escolha dos compostos dentro de um final de semana. A vida do pneu, que representa o número de voltas percorridas por um mesmo jogo de pneus, e o composto utilizado são, portanto, diretamente ligados à física da degradação e figuram entre as variáveis preditoras mais relevantes deste trabalho. Além do desgaste dos pneus, a massa do carro reduz progressivamente ao longo da corrida à medida que o combustível é consumido, tornando o carro mais leve e, consequentemente, mais rápido. Esse efeito, somado à evolução da própria degradação dos pneus, faz com que o tempo de volta apresente uma tendência sequencial dentro de cada *stint*, o que motiva o uso da variável tempo da volta anterior como preditor e também a busca por modelos que capturem essa tendência de série temporal, como a rede LSTM. A Figura 2 apresenta os *stints* durante a corrida do Bahrein em 2025. Através dela, é possível visualizar a alternância entre compostos e a respectiva vida de cada um deles.

As condições climáticas e da pista também desempenham papel central no ritmo de corrida. A temperatura da pista influencia diretamente a aderência dos compostos e a taxa em que os mesmos degradam-se. Para cada composto inclusive, existe uma janela de temperatura ideal de operação: fora dela, o pneu perde performance seja por falta de aderência (frio) ou por superaquecimento (*overheating*). A Figura 3 auxilia no entendimento do comportamento do pneu descrito, através de um composto de pneu arbitrário. Compostos mais duros, como o C1, tem mais dificuldade de chegar na faixa

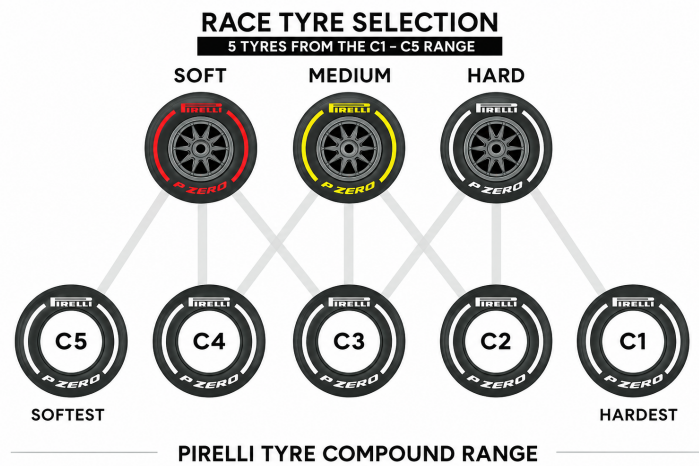


Figura 1 – Exemplificação da escolha de compostos pirelli.

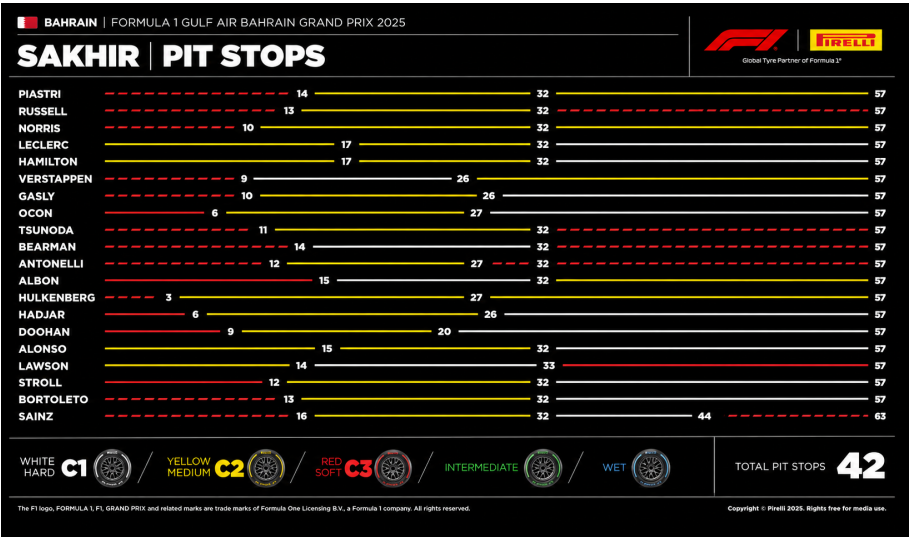


Figura 2 – Stints durante o Grande Prêmio do Bahrein em 2025.

de temperatura ideal, enquanto mais macios, como o C5, rapidamente conseguem atingir sua faixa, em troca da degradação mais acentuada pela sua maciez. A umidade do ar, a pressão atmosférica e a velocidade do vento afetam diretamente o pacote aerodinâmico de cada equipe. Em pistas com longas retas e poucas curvas, equipes almejam pelo *setup* que extraia o máximo possível da sua unidade de potência; dessa maneira, o desempenho nas curvas é prejudicado. O inverso pode acontecer para pistas que tenham curvas de baixa e média velocidade com retas mais curtas. Cada alteração nessas componentes climáticas pode alterar a maneira como as equipes e os pilotos enxergam o *setup* ideal para o final de semana de corrida, fazendo com que os carros tenham um ritmo de corrida diferente em partes do circuito ou até mesmo da corrida. A Figura 4 ajuda a compreender as forças aerodinâmicas que atuam em um carro de Fórmula 1, tais como o *downforce*, o arrasto e a aderência, todas essas sendo impactadas pelo circuito e pela configuração do *setup* do carro.



Figura 3 – Representação do comportamento de um composto de pneu arbitrário durante uma corrida de Fórmula 1.

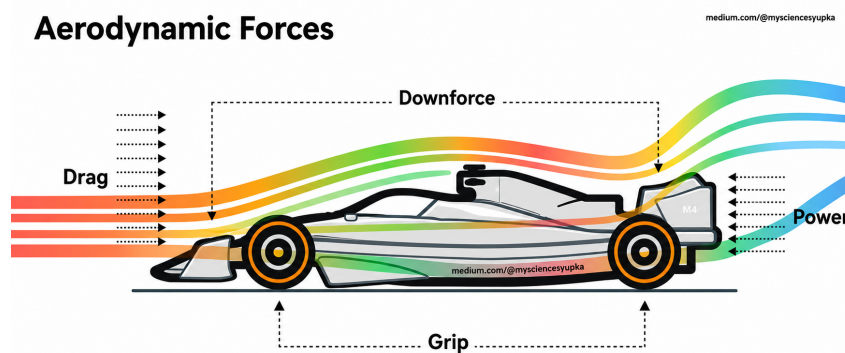


Figura 4 – Representação da aerodinâmica de um carro de F1.

Por fim, como apresentado anteriormente, o ritmo de corrida está sujeito a interrupções decorrentes de eventos aleatórios, como acidentes, entrada de *safety car* (SC) ou *virtual safety car* (VSC) e cenários de chuva, que resultam em voltas significativamente mais lentas e não representativas do ritmo real dos competidores. É justamente a previsão precisa desse ritmo, livre desses eventos atípicos, que permite às equipes avaliar decisões estratégicas, como o momento ideal de realização de uma *pit stop* e a viabilidade de táticas como o *undercut* (realizar a parada antes do rival, aproveitando o pneu novo para recuperar posição) e o *overcut* (estender o *stint* além do rival, aproveitando pista limpa e pneus ainda competitivos). Esses fatores estratégicos, combinados aos fatores físicos apresentados nesta seção, motivam o processamento de dados cauteloso, detalhado posteriormente no trabalho, e reforçam a importância de modelos capazes não apenas de prever o tempo de volta, mas também de explicar quais desses fatores mais o influenciam em cada circuito.

2.3 Circuitos Analisados

Este trabalho analisa cinco circuitos do calendário da Fórmula 1, selecionados de forma a manter a diversidade em termos de tipo de traçado, localização geográfica, período do dia em que a corrida é realizada e condições climáticas típicas. Esta seção apresenta um breve histórico de cada circuito, sua localização e características, justificando o papel de cada um no conjunto experimental deste trabalho.

2.3.1 Circuito Internacional do Bahrein

Inaugurado em 2004, o Circuito Internacional do Bahrein, localizado em Sakhir, foi o primeiro circuito de Fórmula 1 do Oriente Médio. A Figura 5 apresenta o mapa do traçado deste circuito. Construído em meio ao deserto, o traçado é marcado por longas retas seguidas de freadas fortes e por um asfalto altamente abrasivo, que impõe severa degradação aos pneus. Desde 2014, a corrida é realizada no período noturno, sob iluminação artificial, o que reduz a temperatura da pista em relação ao calor extremo do dia no deserto.



Figura 5 – Traçado do Circuito Internacional do Bahrein.

2.3.2 Circuito Corniche de Jeddah

O Circuito Corniche de Jeddah, na Arábia Saudita, estreou no calendário em 2021 e é um dos circuitos de rua mais rápidos do mundo. Localizado à beira do Mar Vermelho, na orla (*corniche*) da cidade de Jeddah, seu traçado combina curvas de alta velocidade em sequência com muros próximos, exigindo precisão extrema dos pilotos. O mapa do traçado deste circuito é apresentado na Figura 6. O circuito une, em uma só pista, velocidade extrema com erro zero de um circuito de rua. Assim como no Bahrein, a corrida é realizada à noite. Por ser um circuito de rua costeiro, o perfil traçado é diferente do encontrado no Bahrein, mesmo que compartilhem regiões geograficamente semelhantes e o mesmo período do dia.

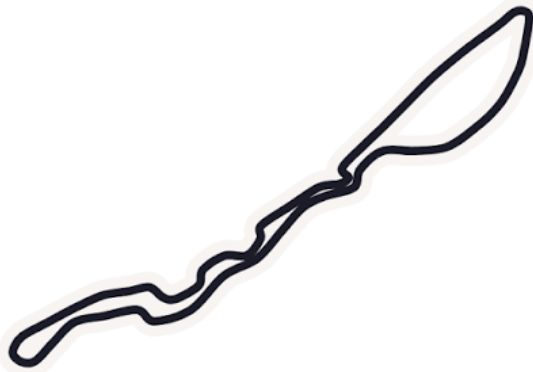


Figura 6 – Traçado do Circuito Corniche de Jeddah.

2.3.3 Circuito das Américas

Inaugurado em 2012 na cidade de Austin, no estado americano do Texas, o Circuito das Américas (do inglês Circuit of the Americas (COTA), foi projetado especificamente para a Fórmula 1, inspirando-se em curvas clássicas de outros circuitos históricos. O mapa do traçado desse circuito é apresentado na Figura 7, esse traçado alterna sequências de curvas de alta velocidade com trechos técnicos de baixa velocidade e uma longa reta oposta. A corrida é realizada no período da tarde e as temperaturas de pista são frequentemente elevadas. Além disso, o asfalto é famosamente conhecido por sua abrasividade e irregularidade. Essa combinação faz do circuito um cenário de forte estresse térmico e mecânico sobre os pneus, em condições diurnas. Dessa maneira, exploramos uma pista em contexto diferente das duas do oriente médio.

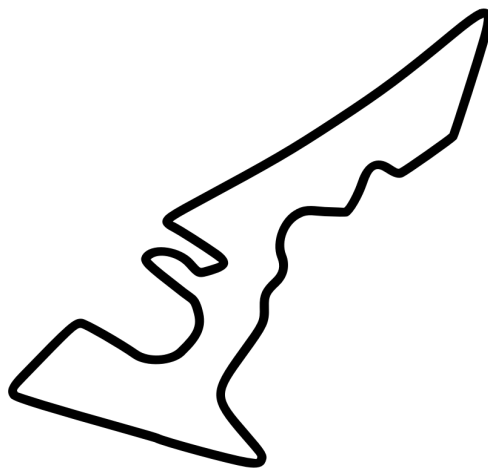


Figura 7 – Traçado do Circuito das Américas.

2.3.4 Circuito Nacional de Monza

A Figura 8 apresenta o traçado do Autódromo Nacional de Monza, na Itália, é um dos circuitos mais tradicionais do automobilismo mundial: inaugurado em 1922, recebe corridas de Fórmula 1 desde a primeira temporada do campeonato, em 1950. Conhecido como o templo da velocidade, é o circuito de maior velocidade média do calendário, caracterizado por longas retas interrompidas por curvas lentas e de frenagem pesada. Por essa natureza, as equipes utilizam em Monza suas configurações de menor carga aerodinâmica do ano, o que torna o desempenho do carro particularmente sensível a variáveis como pressão do ar e vento. A corrida é realizada no período da tarde, no final do verão europeu. Aqui, existe um perfil clássico de pista europeia que as outras três não captavam.

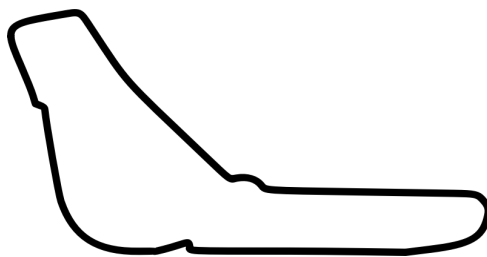


Figura 8 – Traçado do Circuito Nacional de Monza.

2.3.5 Hungaroring

O Hungaroring, localizado em Mogyoród, nos arredores de Budapeste, na Hungria, recebeu em 1986 o primeiro Grande Prêmio de Fórmula 1. Conforme mostra o traçado desse circuito na Figura 9, é um circuito curto, sinuoso e de baixa velocidade média. A sucessão de curvas de baixa e média velocidade, com poucas retas, exige máxima carga aerodinâmica e impõe desgaste contínuo aos pneus, com poucas oportunidades de descanso ao longo da volta. A corrida ocorre no período da tarde, no auge do verão europeu, tipicamente sob temperaturas elevadas. O Hungaroring representa, portanto, o extremo oposto de Monza no espectro de velocidade e carga aerodinâmica.

Em síntese, o conjunto selecionado cobre: dois circuitos noturnos no Oriente Médio com perfis opostos de traçado (deserto abrasivo *versus* circuito de rua costeiro); uma corrida diurna de alto estresse térmico e asfalto abrasivo (Estados Unidos); o extremo de alta velocidade e baixa carga aerodinâmica do calendário (Itália); e o extremo de baixa velocidade e alta carga aerodinâmica (Hungria). Essa diversidade garante que os modelos deste trabalho sejam avaliados sob condições climáticas, físicas e estratégicas substancialmente distintas, sustentando a análise multi-circuito proposta.

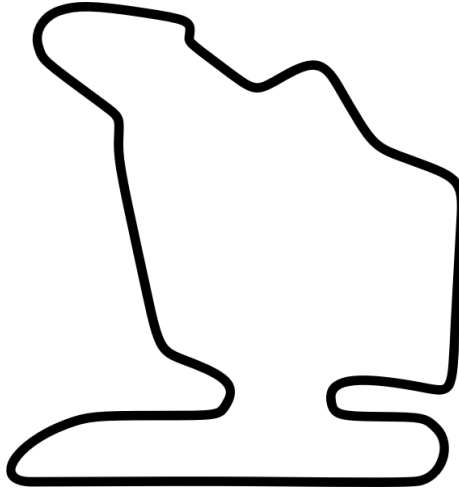


Figura 9 – Traçado do Hungaroring.

2.4 Trabalhos Relacionados

A Fórmula 1 apresenta diversos problemas que podem ser resolvidos através da aplicação de técnicas de aprendizado de máquina e ciência de dados. Historicamente, a previsão de cenários de corrida está diretamente relacionada ao desenvolvimento de simulações que apontam a previsão de determinados cenários, dado um conjunto de escolhas em uma corrida (BEKKER; LOTZ, 2009). No entanto, a previsão desses cenários deve focar em uma área específica, seja ela o ritmo de corrida puro, a quantidade de paradas para troca de pneu, decisões na escolha dos compostos dos pneus ou até mesmo em fatores climáticos e aleatórios, tal como a probabilidade de uma janela de *safety car* originada por uma eventualidade. Dessa forma, a literatura transita entre essas áreas do problema. A Tabela 1 resume os trabalhos relacionados classificados em diferentes tipos de vertentes de pesquisa existentes na literatura e que serão discutidos a seguir.

Em primeiro lugar, existe uma vertente de trabalho de simulação de corridas e otimização de estratégias, em que o foco está na ocorrência de *pit stops* e na escolha dos compostos dos pneus. Nessa vertente, podem ser citadas simulações capazes de representar: (a) ultrapassagens, falhas mecânicas e *pit stops* em um Grande Prêmio (BEKKER; LOTZ, 2009); (b) degradação dos pneus, perda de massa de combustível, paradas e ultrapassagens (HEILMEIER; GRAF; LIENKAMP, 2018); (c) eventos aleatórios como largadas, acidentes, variações de tempos de volta e fases de *safety car* (HEILMEIER et al., 2020); e (d) degradação dos pneus (GARCÍA, 2024). Esses trabalhos mostram que a estratégia na F1 depende de uma boa simulação do ambiente e de estimativas confiáveis do ritmo de corrida.

Outra vertente de pesquisa, diretamente relacionada a este trabalho, utiliza algoritmos clássicos de aprendizado de máquina para prever resultados, desempenho e variáveis associadas à performance em Fórmula 1. Em (CHETELES, 2024), os autores exploram a

importância e seleção de atributos na previsão de posições finais e organizam atributos em grupos relacionados ao clima, carro e piloto. Ainda, o trabalho de (BANSAL et al., 2025) propõe o uso de aprendizado de máquina para prever a performance e a pontuação no campeonato utilizando 70 anos de dados históricos de categoria. De forma semelhante, em (KRZYSZTOŃ; SMOLKA, 2026) é realizada a classificação de resultados de corrida utilizando diferentes algoritmos. O trabalho de (TODD et al., 2025) explora juntamente com a Mercedes-AMG Petronas diferentes modelos em um contexto de degradação de pneus, estimando a energia aplicada sobre eles. Por fim, em (HABER et al., 2025), os autores utilizam *Random Forest* em conjunto com a análise de *Shapley Additive Explanations* (SHAP) para prever o vencedor da corrida, fornecendo também interpretabilidade para o modelo proposto. Essa abordagem se aproxima diretamente do objetivo de extração de *insights* estratégicos almejado por este trabalho, ainda que aplicada a uma variável de saída distinta (posição final, em vez do tempo de volta). Esses trabalhos se concentram, majoritariamente, em previsões de posição final, pontuação ou resultado de corrida, mas a previsão do tempo de volta para cada volta da corrida não é explorada.

Uma terceira vertente complementar à anterior utiliza modelos de aprendizado profundo para capturar padrões complexos em dados de corrida. Em (FATIMA; JOHRENDT, 2023), por exemplo, é proposto o modelo Deep-Racing, baseado em redes neurais profundas com camadas de *embeddings*, voltado à previsão de ranking e estratégia ótima de parada. De forma semelhante, (SASIKUMAR; LEEMA; BALAKRISHNAN, 2025) desenvolve um arcabouço orientado por dados que utiliza diversos modelos de aprendizado profundo para prever o momento ideal de realização de *pit stops*. Por sua vez, em (ZHAO, 2024), os autores utilizam uma rede neural profunda para prever tempos de volta em sessões de classificação, considerando características de pilotos e circuitos. Além de exigirem um maior volume de dados e terem um custo computacional elevado, modelos de aprendizado profundo deixam a desejar no quesito de interpretabilidade. Isso pode limitar sua aplicação quando o objetivo não é apenas prever, mas também compreender quais fatores explicam o ritmo de corrida.

Dentro dessa terceira vertente, uma quarta linha de pesquisa, mais diretamente relacionada a este trabalho, concentra-se especificamente na predição do tempo de volta. Como mencionado, Zhao (2024) utiliza uma rede neural profunda para prever tempos de volta em sessões de classificação, separando a série temporal por circuito e por piloto. O principal problema dessa exigência é que ela reduz os dados disponíveis e pode tornar o modelo menos preciso. De forma semelhante, Kara e Küçükmanisa (2025) prediz tempos de volta utilizando uma rede neural profunda LSTM, mas sem separar a série temporal por circuito. Isso é problemático porque combinar dados de diferentes circuitos em uma única série temporal produz modelos imprecisos, já que cada circuito apresenta condições climáticas significativamente distintas que afetam o tempo de volta. Em suma, além do alto custo computacional, os modelos de aprendizado profundo exigem grandes volumes

de dados para produzir resultados precisos, o que se torna um problema central quando a série temporal precisa ser segmentada por piloto e por circuito.

Recentemente, é natural que a estratégia de corrida venha sendo tratada como um problema sequencial de tomada de decisão (PICCINOTTI, 2021), o que aproxima a Fórmula 1 do campo de aprendizado por reforço enquanto desenvolve agentes inteligentes que tomam decisões em tempo real ao longo da corrida (THOMAS et al., 2025). Apesar disso, a eficácia desses métodos depende da qualidade do simulador, especificidade dos dados e dos modelos utilizados para representar tempos de volta, degradação do pneu, tráfego, ultrapassagens e eventos aleatórios.

Adicionalmente, entre modelos de previsão e tomada de decisão estratégica, existem estudos que integram modelos de aprendizado de máquina à ambientes de simulação. No contexto da Fórmula-E, em (LIU; FOTOUHI, 2020) redes neurais artificiais são combinadas com Monte Carlo Tree Search para desenvolver estratégias de gerenciamento de energia. Esse trabalho demonstra como modelos preditivos funcionam como componentes de sistemas de decisão, o que é particularmente importante para a Fórmula 1, pois esse modelo pode ser incorporado em simulações estocásticas. É útil para testar estratégias de *pit stop* (e.g. *undercut* ou *overcut*) e respostas a eventos inesperados.

Diante desse cenário, apesar de todos esses estudos, ainda existem lacunas na literatura. Poucos trabalhos tratam especificamente da predição do tempo de volta, e, entre os que o fazem, a exigência de segmentar a série temporal por piloto e por circuito reduz consideravelmente o volume de dados disponível, comprometendo a acurácia de modelos que dependem de grandes volumes de dados, como as redes neurais profundas. Além disso, ainda é necessário investigar o ritmo de corrida volta a volta em múltiplos circuitos, combinando desempenho preditivo, validação temporal rigorosa e interpretabilidade, essa última pouco explorada nos trabalhos de predição de tempo de volta revisados, ainda que presente, de forma incipiente, em estudos voltados a outras variáveis de saída (HABER et al., 2025). Este trabalho pretende preencher essa lacuna utilizando dados públicos da biblioteca FastF1, focando no tempo de volta em cinco circuitos distintos, comparando diferentes algoritmos de aprendizado de máquina e propondo um modelo híbrido com aprendizado residual para mitigar o problema da limitação de dados. Dessa forma, o trabalho não apenas busca prever os tempos de volta com maior precisão, mas também extrair *insights* sobre os fatores que impactam o ritmo de corrida em cada pista, estendendo essa análise interpretativa também ao componente não linear do modelo proposto.

Tabela 1 – Resumo dos trabalhos relacionados, organizados segundo as quatro vertentes discutidas na fundamentação teórica.

Simulação de Corridas e Otimização de Estratégias		
Trabalho	Contribuição Principal	Limitação / Diferença
Bekker e Lotz (2009)	Simulação de eventos discretos representando ultrapassagens, falhas mecânicas e <i>pit stops</i> .	Não utiliza aprendizado de máquina para prever desempenho dinâmico.
Heilmeier, Graf e Lienkamp (2018)	Modela degradação de pneus, perda de massa de combustível, paradas e ultrapassagens em uma simulação de corrida.	Coefficientes fixos; não incorpora efeitos probabilísticos.
Heilmeier et al. (2020)	Valida o uso de Monte Carlo para modelar incertezas (<i>Safety Car</i>) e avaliar a robustez das estratégias.	Foca na simulação, mas não usa Machine Learning (ML) para prever o desempenho dinâmico (ex: degradação de pneus).
García (2024)	Desenvolve um modelo de otimização determinístico (MIQP) para achar a estratégia ótima em um cenário perfeito.	Ignora eventos aleatórios, resultando em uma solução que pode ser muito arriscada em um cenário real.
Piccinotti (2021)	Introduz um planejador com Monte Carlo Tree Search (MCTS) para otimização de estratégia.	Exige simuladores dedicados e possui limitações de escalabilidade com a profundidade da busca.
Thomas et al. (2025)	Aplica Aprendizado por Reforço (RL) para otimização de estratégia em tempo real, com foco em explicabilidade.	Altíssima complexidade e custo computacional, sendo pouco acessível para pesquisa independente.
Liu e Fotouhi (2020)	No contexto da Fórmula-E, combina redes neurais com <i>Monte Carlo Tree Search</i> para gerenciamento de energia.	Aplicado a outra categoria (Fórmula-E); trata gerenciamento de energia, não diretamente o <i>pit stop</i> de F1.
Algoritmos Clássicos de ML para Resultados e Desempenho		
Trabalho	Contribuição Principal	Limitação / Diferença
Cheteles (2024)	Explora importância e seleção de atributos (clima, carro, piloto) na predição de posições finais.	Foco em posição final, não em tempo de volta; não trata dados sequenciais volta a volta.
Bansal et al. (2025)	Usa Gradient Boosting para prever o resultado da corrida, mostrando a posição de largada como fator dominante.	Abordagem estática e puramente preditiva; não simula nem otimiza a estratégia para alterar o resultado.
Krzysztoń e Smółka (2026)	Classifica resultados de corrida com SVM, Random Forest e Gradient Boosting, otimizados via Optuna.	Foco em classificação de resultado, não em predição contínua do ritmo de corrida.

Algoritmos Clássicos de ML para Resultados e Desempenho		
Trabalho	Contribuição Principal	Limitação / Diferença
Todd et al. (2025)	Em colaboração com a Mercedes-AMG Petronas, estima a energia aplicada aos pneus a partir de dados privados de telemetria.	Utiliza dados privados de equipe, não replicáveis; não trata diretamente o tempo de volta.
Haber et al. (2025)	Usa Random Forest com análise SHAP para prever o vencedor da corrida, fornecendo interpretabilidade ao modelo.	Interpretabilidade aplicada à posição final, não ao tempo de volta; não trata múltiplos circuitos nem dados limitados.
Jafri (2024)	Usa Redes LSTM para prever resultados, concluindo que prever o tempo de volta não é suficiente para prever o final.	Foco em previsão de resultado. A própria conclusão reforça a necessidade de simular a interação estratégica.
Aprendizado Profundo para Padrões Complexos em Dados de Corrida		
Trabalho	Contribuição Principal	Limitação / Diferença
Fatima e Johrendt (2023)	Propõe o modelo Deep-Racing (EDNN) para prever o ranking final e a volta ótima do <i>pit stop</i> .	Foco em previsão, não em otimização de múltiplos cenários. Alto custo computacional.
Sasikumar, Leema e Balakrishnan (2025)	Propõe um arcabouço orientado por dados, com múltiplos modelos de <i>deep learning</i> , para prever o momento ideal do <i>pit stop</i> .	Foco na decisão de parada, não na predição contínua do tempo de volta como série temporal.
Predição Específica do Tempo de Volta		
Trabalho	Contribuição Principal	Limitação / Diferença
Zhao (2024)	Usa uma Rede Neural Profunda (DNN) para prever o tempo de volta na qualificação, separando a série por circuito e por piloto.	A segmentação por circuito/piloto reduz os dados disponíveis e compromete a acurácia do modelo.
Kara e Küçükmanisa (2025)	Prediz tempos de volta com rede LSTM, sem segmentar a série temporal por circuito.	Combinar dados de circuitos distintos em uma única série produz modelos imprecisos, dado que cada pista tem condições climáticas próprias.

2.5 Considerações Finais

O capítulo fundamenta as escolhas metodológicas relacionadas mais à frente de três maneiras diferentes. A primeira, discutida na Seção 2.1, tratou da estrutura da temporada e do fim de semana de Fórmula 1: a distinção entre treinos livres, classificação e corrida principal auxilia na compreensão geral do *pipeline*, além de justificar a decisão metodológica de utilizar exclusivamente voltas da corrida principal na construção do *dataset*. Como objetivo é o ritmo de corrida, é durante a corrida principal, e não em treinos, *sprints* e classificações, que ele se manifesta. Da mesma forma, a discussão sobre os ciclos do

regulamento técnico da FIA fundamenta o recorte temporal de 2022 a 2025 adotado na coleta de dados. Por esses anos pertencerem a um mesmo ciclo regulamentar, existe uma base técnica comum entre os carros, permitindo que a variável relativa ao ano da corrida capture evolução incremental de desempenho, e não uma ruptura de conceito que inviabilizaria a comparação.

A segunda camada, apresentada na Seção 2.2, decompõe o *race pace* em suas principais componentes: degradação dos pneus, redução de massa por consumo de combustível, condições climáticas e da pista, e eventos aleatórios. Cada um desses mecanismos corresponde diretamente a uma decisão tomada na engenharia de *features* e no pré-processamento de dados descritos no Capítulo 3. A física da degradação e o efeito do combustível justificam a inclusão de variáveis relacionadas à vida do pneu, ao composto utilizado e ao tempo da volta anterior como preditores, a discussão sobre a janela de temperatura ideal de operação de cada composto e sobre o impacto das condições climáticas na configuração aerodinâmica do carro fundamenta a seleção de variáveis relacionadas à temperatura da pista, à umidade, à pressão atmosférica e à velocidade do vento. A caracterização de eventos aleatórios, do tipo *safety car*, *virtual safety car* e chuva, como voltas não representativas do ritmo puro justifica diretamente o processo de filtragem de *outliers* aplicado antes da modelagem. Além disso, a definição de *undercut* e *overcut* como táticas dependentes da previsão precisa do ritmo de corrida reforça a relevância prática do problema de predição tratado neste trabalho, para além de seu interesse estritamente técnico. A terceira camada, apresentada na Seção 2.3, justificou a escolha dos cinco circuitos analisados a partir da diversidade de condições que cada um impõe ao ritmo de corrida. De traçados de alta abrasividade e forte degradação, como Bahrein e Estados Unidos, aos extremos opostos de carga aerodinâmica representados por Monza e Hungaroring. A partir dessa diversidade, nasce a justificativa direta para a abordagem multi-circuito adotada na modelagem, em que cada circuito é tratado como um problema de aprendizado independente.

Por fim, a revisão da literatura apresentada na Seção 2.4 identifica que poucos trabalhos tratam especificamente da predição do tempo de volta, e os que o fazem esbarram na limitação de dados imposta pela segmentação da série temporal por piloto e por circuito. Outra lacuna também é a interpretabilidade, que permanece pouco explorada nesse contexto específico, ainda que presente de forma incipiente em trabalhos voltados a outras variáveis de saída, como a posição final da corrida (HABER et al., 2025). Portanto, essas lacunas motivam a agregação multi-piloto por circuito e o modelo híbrido com aprendizado residual, propostos como estratégia para mitigar a limitação de dados discutida na literatura. Além disso, a extensão da análise interpretativa, incluindo o componente não linear do modelo híbrido, atende à carência de explicabilidade identificada nos trabalhos revisados. O capítulo seguinte detalha como cada um desses fundamentos se traduz, em termos concretos, na coleta, no pré-processamento, na engenharia de *features* e na configuração experimental deste trabalho.

3 Metodologia

Esse capítulo apresenta a descrição da metodologia deste trabalho. A Seção 3.1 apresenta todo o processo de manipulação dos dados, dividindo-se em: extração de dados por meio da API na Seção 3.1.1; pré-processamento dos dados a fim de se obter o ritmo puro de corrida na Seção 3.1.2; e análise exploratória de dados e Engenharia de Features na Seção 3.1.3, que trata da seleção de *features*, criação de novas features e transformações de dados necessárias. A Seção 3.2 apresenta a configuração dos experimentos. A Seção 3.3 se aprofunda nos algoritmos que foram escolhidos como baseline e na abordagem de Deep Learning. Na Seção 3.4, é apresentado o indicador que permite equilibrar o desempenho e a variância do modelo. A Seção 3.5 demonstra as ferramentas que serão utilizadas para a análise interpretativa dos modelos. Por fim, a Seção 3.6 sintetiza as ideias deste capítulo e prepara a discussão para os resultados.

3.1 Preparação dos Dados

Como apresentado pela fundamentação, este trabalho exige uma coleta de dados cuidadosa para atender às questões de pesquisas propostas. Dessa maneira, esta seção apresenta a divisão utilizada durante a preparação dos dados que proporcionou o conjunto de dados utilizado para análise dos modelos.

3.1.1 Extração de Dados

A coleta de dados foi realizada por meio da API FastF1¹. As seguintes pistas foram selecionadas com o objetivo de maximizar a diversidade em termos de tipo de circuito, período do dia, limite de velocidade e temperatura: (i) Circuito Corniche de Jeddah, na cidade de Jeddah, Arábia Saudita, um circuito de rua com corrida noturna; (ii) Circuito Internacional do Bahrein, na cidade de Sakhir, Bahrein, localizado no deserto, com corrida noturna; (iii) Circuito das Américas, na cidade de Austin, Estados Unidos, com corrida na parte da tarde; (iv) Circuito Nacional de Monza, na cidade de Monza, Itália, uma pista de alta velocidade com corrida na parte da tarde; e (v) Hungaroring, na cidade de Mogyoród, Hungria, uma pista de baixa velocidade com corrida na parte da tarde.

Com objetivo de obter um conjunto de dados mais homogêneo e representativo, foram selecionados registros referentes aos anos de 2022 até 2025, uma vez que o regulamento técnico da corrida é o mesmo. Para este trabalho, foram consideradas apenas registros da corrida principal, não foram consideradas corridas *sprint*, treinos livres e classificatórios.

¹ <<https://docs.fastf1.dev/>>

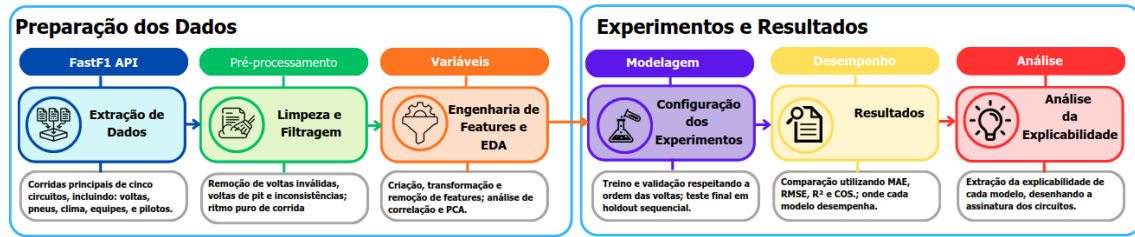


Figura 10 – Pipeline metodológica experimental.

Os dados climáticos registrados ao longo de todo o protocolo de corrida foram utilizados, para cada volta dada, buscou-se um registro climático para mais ou menos trinta segundos após a conclusão da volta, unificando todas as informações no conjunto de dados. Além disso, os nomes das equipes foram padronizados, deixando apenas os dez nomes atuais que terminaram a temporada de 2025. Portanto, equipes como a AlphaTauri e a Alfa Romeo foram atualizadas para os seus respectivos nomes atuais, a saber Racing Bulls e Kick Sauber, respectivamente.

A API da FastF1 trata os compostos de pneus pirelli como *soft*, *medium* e *hard* para todo evento. Sabe-se que a gama de pneus da Pirelli varia de C1 a C5, com o composto C1 representando o mais duro e o C5 o mais macio, logo, o ideal é fazer o mapeamento desses pneus para seu verdadeiro composto. Para isso, foi realizado um *Web scraping* no site da Pirelli² para identificar quais compostos foram utilizados em cada uma das corridas. Dessa forma, isso garante a representação adequada da evolução dos pneus e de seu impacto no tempo de volta na pista. Também evita problemas com os registros, por exemplo, em 2024 nos Estados Unidos foram utilizados os compostos C4 (*soft*), o C3 (*medium*) e o C2 (*hard*), mas em 2025 o composto *hard* se tornou o C1, enquanto que os demais foram mantidos.

A Tabela 2 apresenta o dicionário de dados do dataset. Como pode ser observado, existem nove variáveis numéricas relacionadas à performance, tais como, por exemplo, o tempo de volta (`LapTime_sec`), o tempo de volta anterior (`LapTime_prev`) e a idade do pneu (`TyreLife`). Também existem três variáveis categóricas: (1) `Driver`, que representa o piloto (existem cerca de 30); (2) `Team`, que representa a equipe e (3) `pirelliCompound`, que representa o composto da Pirelli. A variável predita (alvo) é `LapTime_sec`, as demais são variáveis preditoras.

3.1.2 Pré-processamento de Dados

Diferentemente do ritmo puro de corrida, eventos climáticos (e.g., chuva), *safety car*, *virtual safety car* e entradas e saídas de *pit* resultam em voltas significativamente mais lentas que o normal, sendo consideradas registros *outliers*. Dessa forma, um processo de filtragem foi aplicado para remover esses *outliers* e erros do sistema de telemetria.

² <<https://press.pirelli.com/>>

Tabela 2 – Dicionário de dados do dataset.

Variável	Descrição	Tipo	Exemplo
LapTime_sec	Tempo de volta	Numérica	92.45 s
LapTime_prev	Tempo de volta anterior	Numérica	91.80 s
TrackTemp	Temperatura da pista	Numérica	27.80°C
Humidity	Umidade do ar	Numérica	55%
Pressure	Pressão atmosférica	Numérica	992.30 mbar
LapNumber	Número da volta	Numérica	44
TyreLife	Idade do pneu	Numérica	12
Year	Ano da corrida	Numérica	2025
Team	Equipe	Categórica	Team_Haas
Driver	Piloto	Categórica	Driver_GAS
pirelliCompound	Composto de pneu (C1 a C5)	Categórica	pirelliCompound_C3
WindSpeed	Velocidade do vento	Numérica	2.30 m/s

Tabela 3 – Quantidade de voltas registradas por pista, antes e após a filtragem dos dados.

Pista	Dados brutos	Dados filtrados	Retenção (%)
Bahrein	4438	3161	71.2
Arábia Saudita	3562	2660	74.7
Hungria	5358	4213	78.6
Itália	3912	2994	76.5
Estados Unidos	4132	3063	74.1
Total	21402	16091	75.2

Além disso, o 95º percentil foi utilizado como limiar de corte, uma vez que inspeções manuais revelaram diversos *outliers* nas caudas da distribuição de cada *feature*. Assim, o número de instâncias no conjunto de dados é: (a) 3.161 no Bahrein; (b) 2.660 na Arábia Saudita; (c) 3.063 nos Estados Unidos; (d) 2.994 na Itália; (e) 4.213 na Hungria. A Tabela 3 apresenta a quantidade dados brutos *versus* a quantidade de dados filtrados. Além disso, a tabela conta com a coluna de retenção, que mostra qual a porcentagem de voltas que permaneceram após a filtragem proposta. É possível constatar que, em média, pouco mais de 70% das voltas permanecem após o filtro. Por meio das tabelas de sumarização estatística apresentadas no Apêndice A, é possível compreender o comportamento dos dados em cada uma das pistas antes e após a filtragem realizada.

Os valores ausentes nas *features* numéricas foram imputados por meio da mediana. Além disso, a coleta original de dados inclui *features* categóricas nominais: identificadores de piloto (*Driver*), identificadores de equipe (*Team*) e identificadores de composto de pneu (*pirelliCompound*). A técnica de One-hot Encoding codifica essas *features* em dados numéricos. Os dados foram normalizados para apresentar média zero e desvio padrão unitário. A variável alvo, *LapTime_sec*, foi mantida em segundos, sem transformação ou escalonamento.

3.1.3 Engenharia de Features e Análise Exploratória de Dados

A relação sutil entre a temperatura do ar (**AirTemp**) e a temperatura da pista (**TrackTemp**) motivou a criação da variável **TempDelta**, definida como

$$\Delta T = T_{TrackTemp} - T_{AirTemp}, \quad (3.1)$$

em que $T_{TrackTemp}$ representa a temperatura da pista e $T_{AirTemp}$ representa a temperatura do ar. De acordo com a física, $T_{TrackTemp}$ é o preditor primário da aderência mecânica e da degradação térmica do pneu, enquanto ΔT indica o balanço de energia na superfície (ADWAN et al., 2021). Um ΔT elevado indica forte radiação e aquecimento ativo do asfalto, enquanto um ΔT baixo sugere redução da radiação por cobertura de nuvens ou resfriamento.

A análise de correlação, demonstrada na Figura 11, foi realizada por meio do coeficiente de correlação de Pearson, definido como ρ . Para evitar problemas de multicolinearidade, pares de *features* fortemente correlacionadas, i.e. $\rho > 0.85$, foram identificadas como candidatas à remoção, uma vez que a presença de duas ou mais variáveis independentes com correlação alta compartilham informação redundante. No entanto, remover uma variável de um par correlacionado exige um critério. Por esse motivo, a decisão de qual variável do par manter foi tomada com base na Análise de Componentes Principais, do inglês, Principal Component Analysis (PCA), mais especificamente na análise de seus *loadings*.

A PCA transforma o conjunto original de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas entre si, chamadas de componentes principais, do inglês, Principal Components (PC). Os componentes principais são ordenados de forma decrescente pela quantidade de variância explicada dos dados originais que cada componente consegue explicar. Cada componente principal é constituído por uma combinação linear das variáveis originais, e o *loading* de uma variável em um componente representa o peso, ou coeficiente, dessa variável nessa combinação linear. Portanto, quanto maior o valor absoluto do *loading* de uma variável em um componente principal de alta variância explicada, maior a contribuição dessa variável para a estrutura de variação predominante nos dados daquele circuito. Assim, quando duas variáveis apresentam correlação forte entre si, comparar seus *loadings* nos primeiros componentes principais permite identificar qual delas carrega mais informação sobre o comportamento dos dados. Dessa maneira, a variável que vai permanecer pode ser selecionada. A Figura 12 apresenta os gráficos de *loadings* para os componentes PC1 e PC2 de cada circuito analisado ao qual uma variável foi removida. O eixo x diz respeito à variância que a variável explica no PC1 e o eixo y do PC2.

Esse critério foi aplicado a três circuitos, em que pares de variáveis com $\rho > 0.85$ foram identificados. Conforme mostra a Figura 12(a), para a pista dos Estados Unidos, três pares de variáveis apresentaram correlação forte entre si: **TrackTemp** e **TempDelta**

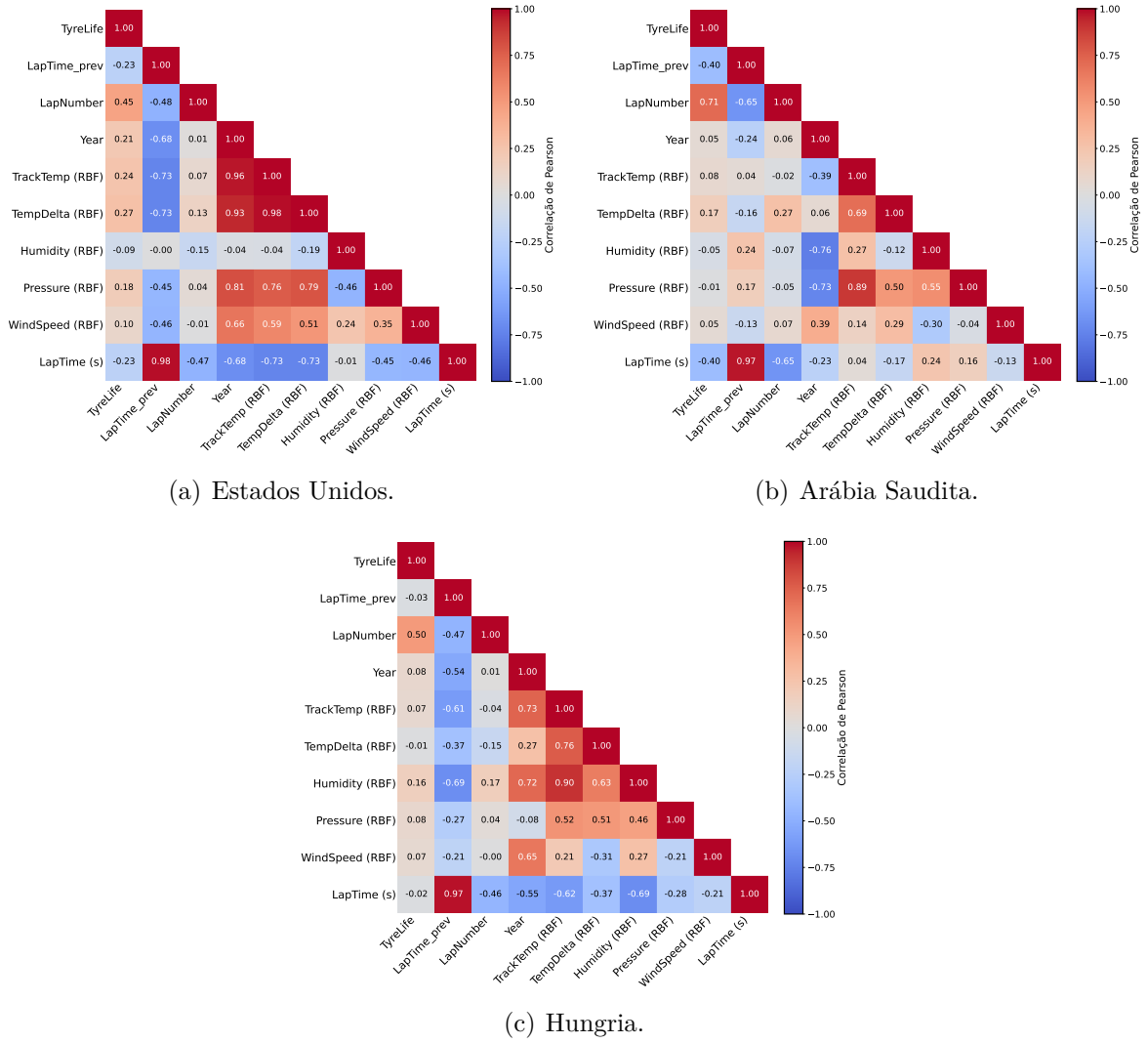


Figura 11 – Matriz de correlação.

($\rho = 0.98$), **Year** e **TrackTemp** ($\rho = 0.97$), e **Year** e **TempDelta** ($\rho = 0.94$). A análise dos *loadings* mostrou que as três variáveis apresentam magnitude de contribuição semelhante nos componentes PC1 e PC2, o que tornaria a escolha, em princípio, arbitrária do ponto de vista puramente estatístico. Nesse caso, a decisão de manter **TrackTemp** e remover **Year** e **TempDelta** foi complementada por um critério físico: o Circuito das Américas está localizado no Texas, região conhecida por um asfalto particularmente abrasivo, no qual a temperatura da pista desempenha papel especialmente relevante na degradação dos pneus conforme discutido na Seção 2.2.

Já na Arábia Saudita, o par **Pressure** e **TrackTemp** apresentou $\rho = 0.89$. Diferentemente do caso anterior, a análise de *loadings* revelou uma diferença clara entre as duas variáveis conforme é exibido na Figura 12(b): **Pressure** apresentou maior contribuição no PC2 e uma contribuição semelhante no PC1, sendo portanto a variável selecionada, enquanto **TrackTemp** foi removida do conjunto de *features* para esse circuito. Por fim, na Hungria, o par **TrackTemp** e **Humidity** apresentou $\rho = 0.89$. Nesse caso, as duas variáveis

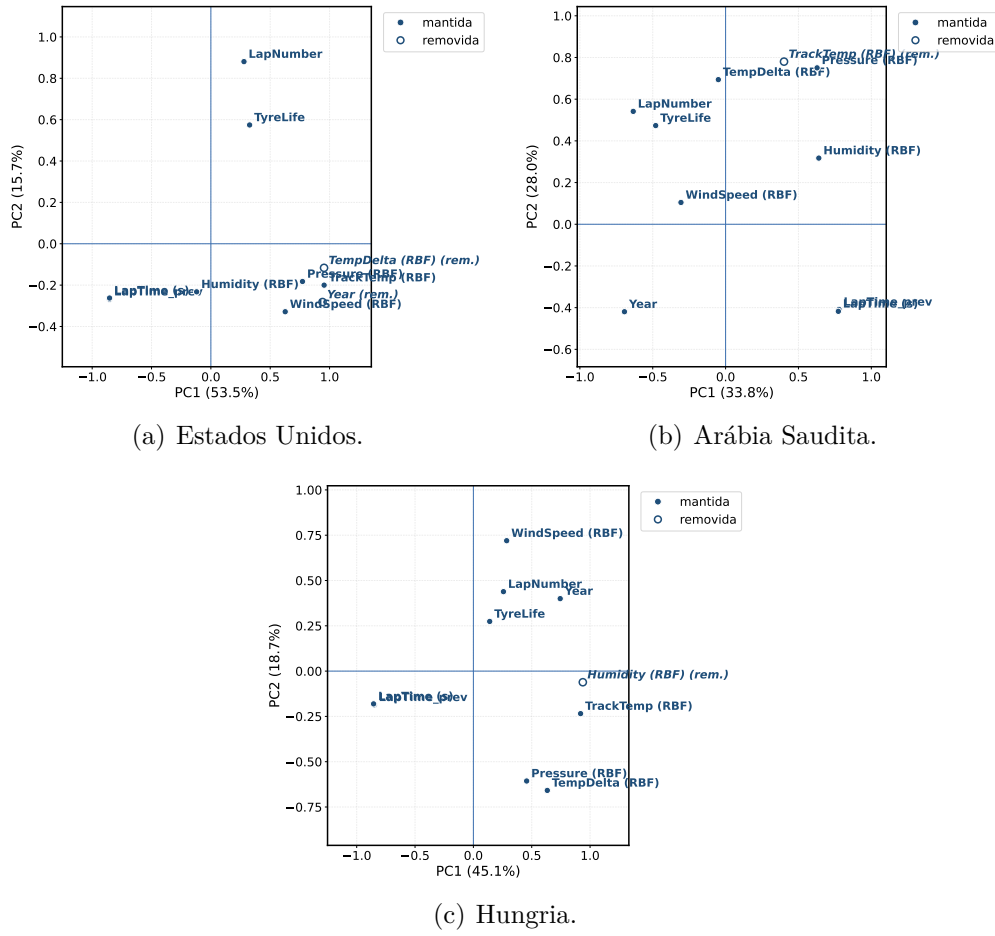


Figura 12 – Loadings PCA.

mostraram contribuição semelhante em PC1, mas **TrackTemp** apresentou *loading* de magnitude superior em PC2 (vide Figura 12(c)), o que motivou sua manutenção e a remoção de **Humidity**. Para os demais circuitos, Bahrein e Itália, nenhum par de variáveis ultrapassou o limiar de $\rho > 0.85$, de modo que todas as *features* originais foram mantidas.

Por fim, cabe destacar que a variável preditora **LapTime_prev** está fortemente correlacionada com a variável alvo **LapTime_sec** em todos os cinco circuitos ($\rho \geq 0.96$). Diferentemente dos casos anteriores, essa correlação não é tratada como um problema de multicolinearidade a ser resolvido pela remoção de uma das variáveis, uma vez que **LapTime_prev** é justamente a variável que captura a forte dependência temporal do tempo de volta discutida na Seção 2.2, e sua manutenção como preditor é uma decisão metodológica central deste trabalho, e não um efeito colateral indesejado da engenharia de *features*.

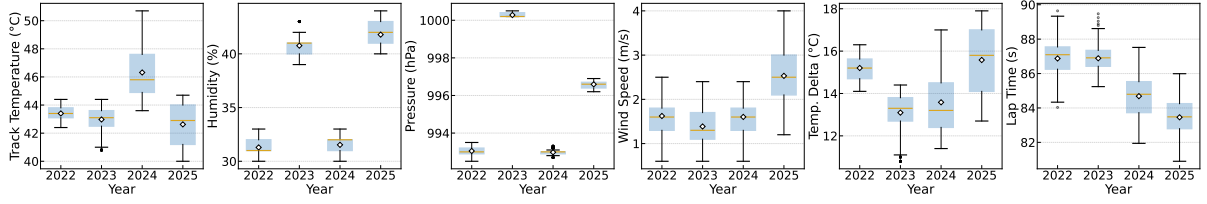
Na Seção A.1 do Apêndice A são apresentadas as tabelas de sumarização estatística dos dados para todas as pistas após o pré-processamento, filtragem e engenharia de *features*. Nessas tabelas, são apresentados as seguintes medidas de localização, dispersão e forma: $\bar{x}$ que representa média; s representa o desvio padrão; CV é o coeficiente de variação; $Q1$

denota o primeiro quartil; $Q2$ é o segundo quartil; $Q3$ representa o terceiro quartil; γ_1 representa o coeficiente de assimetria; e, por fim, κ é o coeficiente de curtose.

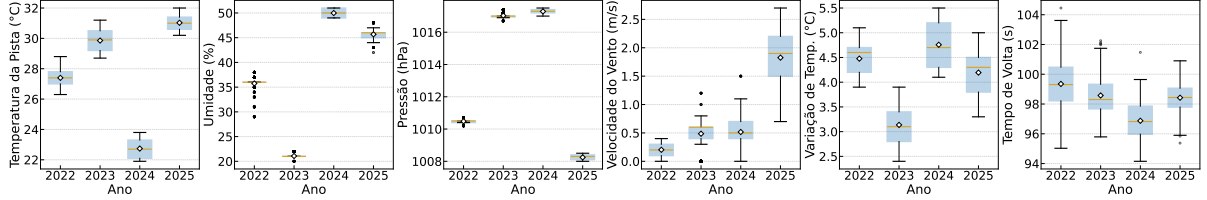
Para facilitar a leitura e interpretação das medidas de sumarização estatística, será utilizado apenas um exemplo de interpretação da sumarização estatística. Para isso, a pista do Bahrein será utilizada e sua respectiva tabela de sumarização estatística é apresentada na Tabela 10. Em geral, a média amostral de todas as *features* é bem próxima de $Q2$ e o $CV \leq 1$ indica menor variabilidade com a maioria dos valores concentrados próximos à média, o que pode ser observado para a maioria das *features* com um intervalo interquartil $Q3 - Q1$ mais estreito. Para *feature WindSpeed*, $\bar{x} > Q2$ pode indicar uma distribuição *right skewed*, o que poderia ser confirmado pelo $\gamma_1 > 1$ destacado em azul. No entanto, inspeções visuais da distribuição mostram que *WindSpeed* apresenta distribuição multimodal. Depois do processo de filtragem, para identificar possíveis *outliers*, são utilizados os coeficientes de skewness (γ_1) e de kurtosis (κ). Valores negativos de κ são destacados em vermelho, que significa que os valores das *features* são distribuídos mais distantes da média e a probabilidade de *outliers* nas caudas da distribuição é menor, uma vez que a distribuição é menos assimétrica, como mostra seus respectivos $\gamma_1 < 1$. Por outro lado, valores de κ positivos podem ser um indicativo da presença de *outliers*, para confirmar, é necessário observar se $\gamma_1 > 1$, mas isso não ocorre em nenhuma das *features*, o que mostra que não existem *outliers*.

Como apresentado no Capítulo 2, a natureza da Fórmula 1 se aproxima mais de uma série temporal por piloto, ano e circuito do que uma série temporal total que envolva apenas o circuito. Esse conceito ganha força ao observar a natureza do tempo de volta para cada piloto. Na Seção A.3 do Apêndice A, são apresentadas as tabelas que apresentam a média do tempo de volta por piloto e por ano, e também a média por equipe e por ano. Em todos os circuitos, é possível observar diferenças claras na média dos tempos de volta de cada um dos pilotos e equipes. Principalmente no âmbito do piloto, essa informação reforça a ideia que cada sequência de tempo de volta dos pilotos em uma determinada corrida constitui uma série temporal única. Além disso, os anos também impactam muito, principalmente pela evolução dos carros, condições climáticas distintas e a introdução ou saída de pilotos.

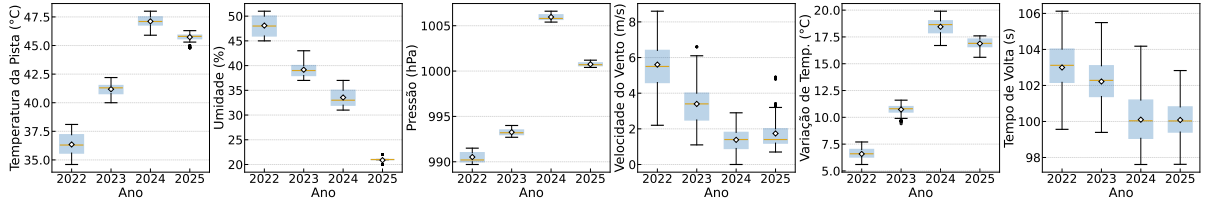
Para enfatizar as características específicas de cada circuito com base nas suas características específicas de sua localização geográfica, a Figura 13 apresenta a distribuição das *features* de condições climáticas e da variável dependente (tempo de volta) para todos os circuitos. Essa figura ilustra que essas variáveis diferem substancialmente entre os anos. Os gráficos de densidade e histogramas foram omitidos para se manter a clareza do texto, no entanto, os principais insights sobre a forma das distribuições das *features* também são exibidos nas tabelas de sumarização estatística presentes na Seção A.1 do Apêndice A. Após inspeções manuais das distribuições das *features*, constatou-se que, embora as variáveis



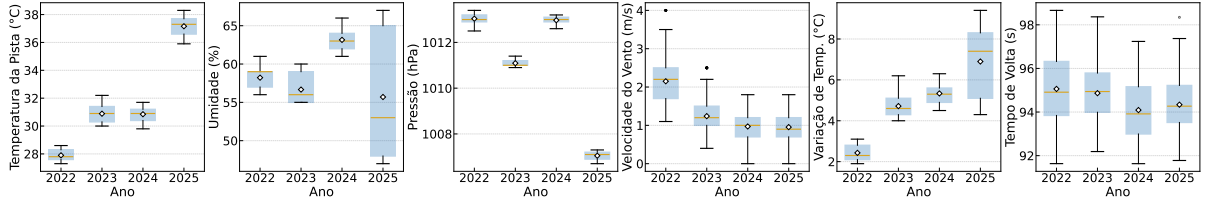
(a) Circuito Nacional de Monza na Itália.



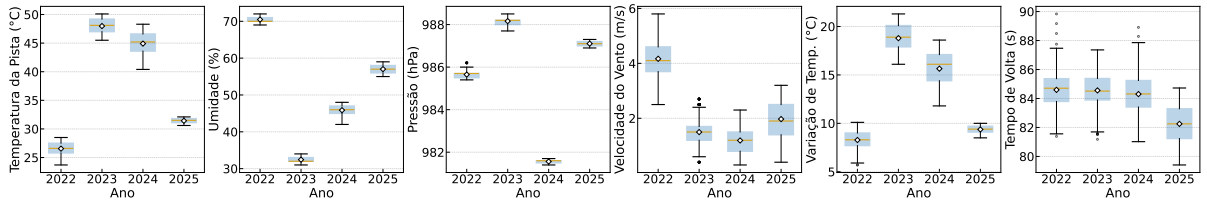
(b) Circuito Internacional do Bahrein no Bahrein.



(c) Circuito das Américas nos Estados Unidos.



(d) Circuito Corniche de Jeddah na Arábia Saudita.



(e) Circuito de Hungaroring na Hungria.

Figura 13 – Distribuição das *features* de condições climáticas e do tempo de volta.

climáticas não sejam *right-* ou *left-skewed*, elas apresentam distribuição multimodal para todas as pistas (i.e., *Humidity*, *Pressure*, *TrackTemp*, *WindSpeed* e *TempDelta*). Portanto, para essas *features*, foi aplicada a transformação Gaussian Radial Basis Function (RBF) Kernel, utilizando uma medida de similaridade em relação à respectiva mediana, que serve como ponto de referência (GÉRON, 2019).

Adicionalmente, para enriquecer o conjunto de dados, também foi realizado um *Webscrapping* para obter os compostos reais de pneu utilizados durante as corridas. Essa abordagem visa preencher lacunas que podem existir na generalização dos compostos.

Os três compostos que são escolhidos podem mudar e seria impossível capturar todas as nuances se o composto C3 fosse tratado como C2. A Figura 14 exemplifica o comportamento dos compostos de pneu ao longo dos anos para cada uma das pistas. Para a Itália, o C3 e o C4 são os mais utilizados, mas em alguns anos, o C4 era o mais macio. Em 2025, por exemplo, ele era o composto intermediário, e o C5 seria o composto mais macio. Mesmo assim, o C3 e o C4, que apresentavam nomenclaturas diferentes nas corridas do final de semana, continuaram sendo os mais utilizados. Para os Estados Unidos, a situação é ainda mais diferente, em 2025 existe um salto entre os compostos, com a categoria utilizando C1, C3 e o C4, pela primeira vez em qualquer ano registrado. O composto C2, que era muito utilizado como o pneu médio, dessa vez é descartado. Esses detalhes tornam esse trabalho extra extremamente importante para a análise deste trabalho.

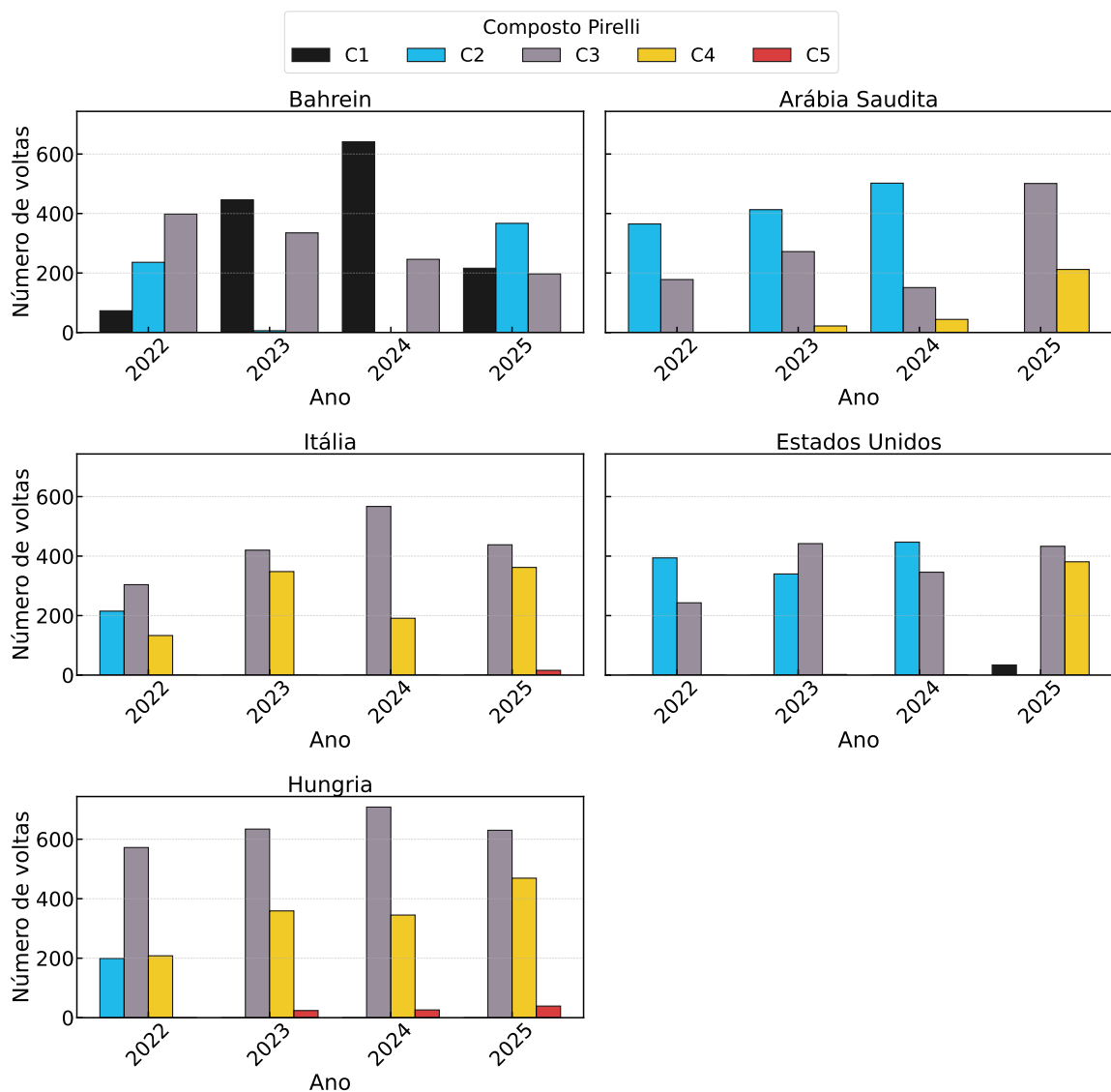
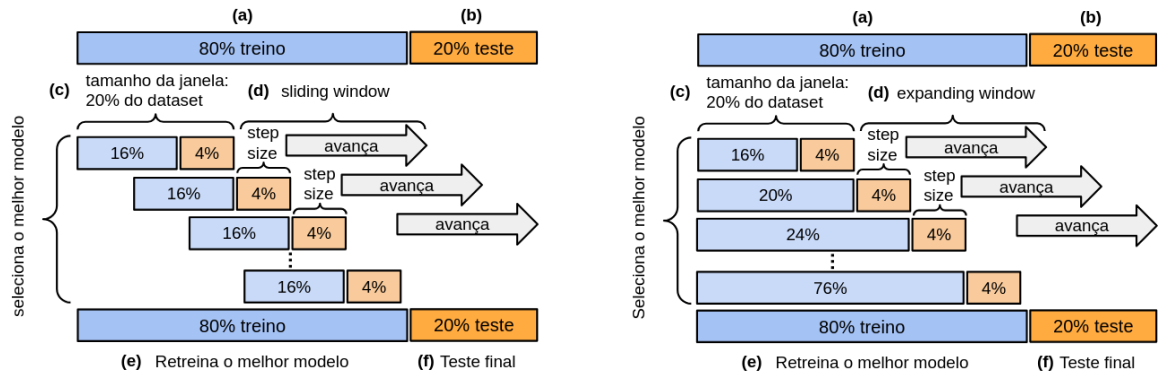


Figura 14 – Número de voltas por composto de pneu e ano em cada GP.

(a) Exemplo de *sliding window validation*.(b) Exemplo de *expanding window validation*.Figura 15 – Detalhes das configurações dos experimentos *baselines*.

3.2 Configuração dos Experimentos

Para avaliar a capacidade de generalização dos modelos, foram utilizadas duas estratégias de validação baseadas no *rolling forward/forecast* para séries temporais (HYNDMAN; ATHANASOPOULOS, 2021). A primeira foi a validação por janela deslizante, do inglês, Sliding Window (SW), e a segunda, a validação por janela expansiva, do inglês, Expanding Window (EW), que mantém a ordem cronológica das voltas das corridas (PEIXEIRO, 2022). Essa estratégia evita que voltas futuras influenciem o treinamento, i.e. *data leakage*, o que seria comum se fosse utilizada uma validação cruzada convencional. O dataset, ordenado por volta e ano, foi dividido em 80% para treino e validação e 20% para o teste. A Figura 15(a) ilustra o funcionamento do SW em que a janela de treinamento possui tamanho fixo (*window size*) e um *step size* de deslocamento para frente. As janelas do SW são divididas sempre em: 80% treinamento e 20% validação. Por sua vez, o EW é representado na Figura 15(b), na qual a janela começa a partir de uma *window size*, dividido também dividida em 80% para treino e 20% para teste, e expande a cada *fold* de acordo com um *step size* de deslocamento para frente que aumenta o conjunto de treino na medida em que avança. Nessa abordagem, o modelo acumula todo o histórico de voltas da corrida, considerando um histórico de voltas de longo prazo, enquanto que um histórico de voltas de curto prazo é considerado pelo SW. Para definir o tamanho da janela, foram realizados testes com tamanhos entre 5% e 50% do total de instâncias do *dataset* em cada circuito e para cada modelo.

3.3 Algoritmos Baseline e Abordagem de Aprendizado Profundo em Dois Estágios

Para os experimentos, a Regressão Linear Múltipla e o algoritmo de aprendizado por *ensemble* XGBoost foram utilizados como *baselines*. O XGBoost foi selecionado por

Tabela 4 – Espaço de busca do Optuna para o XGBoost-EW por pista.

Hiperparâmetro	Bahrein	Arábia S.	Hungria	Itália	EUA
learning_rate	[0.0067, 0.3]	[0.0067, 0.3]	[0.0067, 0.3]	[0.0167, 0.3]	[0.0167, 0.5]
max_depth	[3, 7]	[1, 7]	[1, 5]	[1, 7]	[3, 9]
min_child_weight	[4, 6]	[4, 6]	[4, 6]	[4, 6]	[1, 6]
subsample	[0.656, 0.844]	[0.678, 0.872]	[0.678, 0.872]	[0.678, 0.872]	[0.719, 1]
colsample_bytree	[0.744, 0.956]	[0.722, 0.928]	[0.722, 0.928]	[0.722, 0.928]	[0.809, 1]
gamma	[0, 0.7]	[0, 0.7]	[0, 0.7]	[0, 0.7]	[0, 0.7]
reg_alpha	[3.33e-04, 0.03]	[3.33e-04, 0.03]	[3.33e-04, 0.003]	[3.33e-04, 0.03]	[3.33e-06, 0.03]
reg_lambda	[0.167, 3]	[0.167, 3]	[0.167, 3]	[0.167, 3]	[0.0333, 3]

Tabela 5 – Hiperparâmetros selecionados para o XGBoost-EW (mediana entre folds) por pista.

Hiperparâmetro	Bahrein	Arábia Saudita	Hungria	Itália	Estados Unidos
learning_rate	0.165	0.1993	0.0429	0.0619	0.2015
max_depth	3	3	1	1	6
min_child_weight	5	4	6	5	4
subsample	0.6842	0.7561	0.7693	0.7846	0.7841
colsample_bytree	0.9042	0.8332	0.9097	0.7794	0.8876
gamma	0.2035	0.3163	0.5047	0.1484	0.213
reg_alpha	0.0014	0.0032	0.0012	0.0026	6.24e-04
reg_lambda	0.5102	0.2834	0.5312	0.265	0.4259
n_estimators	249	152	1050	395	85

sua eficiência em capturar relações não lineares e interações complexas entre variáveis. Para o algoritmo XGBoost, o ajuste de hiperparâmetros foi realizado por meio da biblioteca Optuna³, um *framework* de otimização de hiperparâmetros. Os hiperparâmetros incluíram a taxa de aprendizado, a profundidade máxima da árvore, a soma mínima dos pesos das instâncias em um nó folha, a proporção de *subsample* das instâncias de treino, a proporção de *subsample* de colunas, a redução mínima de perda (*gamma*) e os parâmetros de regularização L1 e L2. Inicialmente, para configurar o espaço de busca, alguns valores de hiperparâmetros foram propostos e executados em 80% dos dados de treino e em 20% dos dados de teste. Com base nos resultados dessa configuração inicial, um espaço de busca personalizado foi criado para cada circuito. A Tabela 4 apresenta os valores definidos para o espaço de busca de cada circuito, seguindo a abordagem EW. Para explorar diferentes combinações de hiperparâmetros do modelo dentro desse espaço de busca, foram definidas 100 *trials*. Para cada *trial*, o modelo é treinado com *early stopping* nas voltas de treino e testado nas voltas de validação de cada circuito. O modelo final é treinado em 80% do conjunto de treino. Os hiperparâmetros são selecionados por meio da mediana dos melhores hiperparâmetros de cada janela. O modelo é então testado nos 20% do conjunto de teste. Os hiperparâmetros dos modelos finais EW são apresentados pela Tabela 5.

O protocolo de validação final utilizado para os *baselines* foi o EW. Durante a configuração dos *baselines*, o *expanding window validation* se destacou como a estratégia

³ <<https://optuna.org/>>

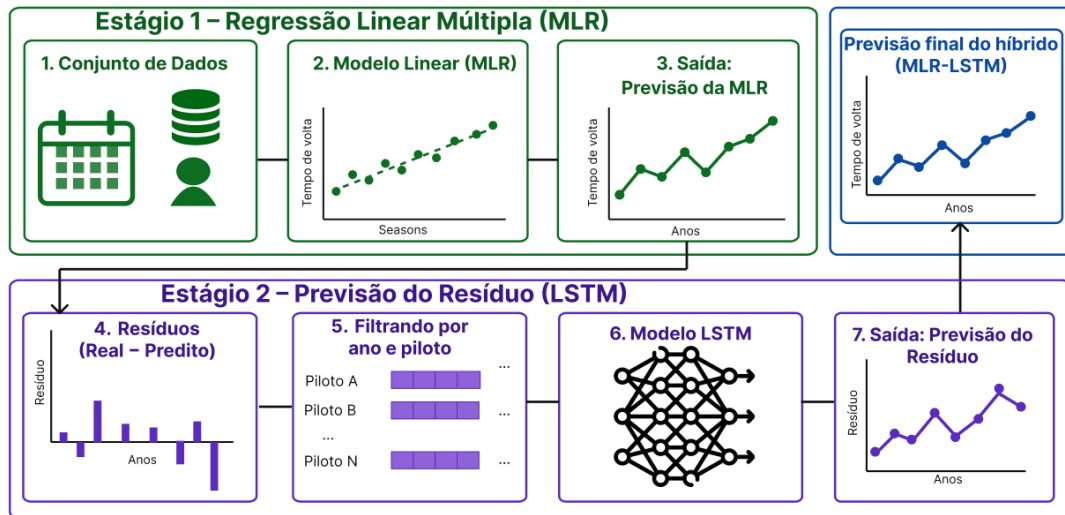


Figura 16 – Abordagem Híbrida de Dois Estágios.

mais robusta, apresentando melhor desempenho na Regressão Linear e por boas margens no XGBoost. Por essa razão, a partir daqui, os resultados apresentados, tais como o espaço de busca e os hiperparâmetros das Tabelas 4 e 5, são exclusivamente da janela expansiva⁴. Em circuitos como Hungria e Bahrein, o *sliding window validation* obteve resultados semelhantes em janelas em torno de 30%, mas a consistência entre treino e teste, observada por meio da métrica de COS foi determinante em favor do EW.

Modelos de aprendizado profundo exigem grandes volumes de dados para aplicações de alto desempenho. Por esse motivo, e seguindo Fong et al. (2020), são utilizados neste trabalho modelos híbridos para melhorar a acurácia e fornecer previsões melhores com conjuntos de dados limitados. A estratégia consiste em combinar a regressão linear tradicional, que produz previsões lineares, com uma rede neural profunda do LSTM, que captura dependências temporais não lineares em dados sequenciais. Essa combinação é chamada de MLR-LSTM híbrido, que utiliza uma abordagem de aprendizado residual com previsão em dois estágios (VATSA et al., 2024; MAHMUD et al., 2025). Essa estratégia é apresentada na Figura 16. Primeiramente, o componente MLR modela as tendências de longo prazo do ritmo de corrida ao longo de múltiplas temporadas, utilizando os tempos de volta agregados de todos os pilotos em um determinado circuito. Em seguida, as diferenças entre os valores reais e os valores preditos, ou seja, os resíduos, são extraídas e tratadas como uma nova série temporal, filtrada por ano para cada piloto. Assim, o LSTM modela as variações aleatórias desses resíduos do MLR como uma série temporal multivariada, utilizando a série temporal das *features* como entrada. Ao final, a previsão do modelo híbrido é obtida somando-se as previsões do MLR aos resíduos preditos pelo LSTM. Detalhes dos experimentos com o LSTM são descritos a seguir.

⁴ Por questões de legibilidade e clareza do texto, os resultados exaustivos dos experimentos utilizando EW e SW não são exibidos.

Tabela 6 – Hiperparâmetros selecionados para o LSTM, por pista.

Hiperparâmetro	Bahrein	Arábia Saudita	Hungria	Itália	Estados Unidos
window_ratio	0.3	0.03	0.03	0.05	0.1
sequence_length	17	2	3	3	6
lstm_units	24	16	16	24	16
dense_units	48	24	24	24	48
dropout	0.4132	0.3308	0.3497	0.3779	0.2409
recurrent_dropout	0.0829	0.09	0.0998	0.1432	0.1745
learning_rate	7.18e-04	0.0012	0.0015	0.0014	0.0017
batch_size	16	16	32	32	16
l2_reg	6.93e-04	4.23e-04	7.32e-04	9.54e-04	0.0015
huber_delta	1.5537	1.5148	1.8024	1.4503	1.9851
epochs	11	2	20	2	3

Para variáveis categóricas de alta cardinalidade, como **Driver** e **Team**, foram utilizados *embeddings* para mapear cada piloto e cada equipe em um vetor denso de baixa dimensionalidade, aprendido durante o treinamento (FATIMA; JOHRENDT, 2023). Para esse fim, empregou-se a heurística

$$\min(d_{\max}, (\text{card} + 1)/2), \quad (3.2)$$

em que $d_{\max}$ é a dimensionalidade máxima e card é a cardinalidade das *features* (FATIMA; JOHRENDT, 2023). O ajuste de hiperparâmetros do LSTM foi realizado por meio da biblioteca Optuna⁵, incluindo: o número de unidades na camada recorrente, o número de unidades na camada densa, as taxas de *dropout* e *recurrent dropout*, a taxa de aprendizado, o tamanho do *batch*, o parâmetro *delta* da função de perda Huber, e a regularização L2. Para explorar diferentes combinações dentro do espaço de busca, foram executadas 50 *trials* por circuito. O espaço de busca foi definido com base em execuções iniciais de *baseline* e mantido consistente entre todos os circuitos. O conjunto de treino (80%) é dividido sequencialmente em subconjuntos de treino e validação. O número de épocas é ajustado por meio de *early stopping* com base na perda de validação, combinado com redução adaptativa da taxa de aprendizado. O modelo final é retreinado no conjunto de treino completo (80%) e avaliado nos 20% restantes do conjunto de teste. A Tabela 6 apresenta os valores para os hiperparâmetros do componentes LSTM do modelo híbrido.

O LSTM não utiliza SW ou EW. O *time step* é utilizado para formar uma sequência que reúne as voltas anteriores à volta alvo. O agrupamento dessa sequência é feito por **Year** e **Driver**, de modo que a evolução temporal das voltas de cada piloto seja adequadamente representada durante aquela corrida. Dessa maneira, é impossível que sequências sejam formadas com premissas artificiais, como a quebra de um ano para o outro, e vazamento de informação. Além disso, essa sequência natural das voltas de um mesmo piloto impede que o modelo se baseie demais em informações muito parecidas, como por exemplo: o

⁵ Maiores detalhes sobre o espaço de busca de hiperparâmetros estão disponíveis no repositório de código deste trabalho.

alvo é a volta 20 e o LSTM fornece dez voltas 19 de dez pilotos diferentes. A capacidade de o *time step* conseguir ir além e reconstruir a série temporal daquele piloto se torna interessante para esse tipo de modelo. Para se definir o tamanho ideal dessa sequência, foram testados os seguintes valores de tamanho de janela: 1%, 3%, 5%, 10%, 15%, 20%, 25% e 30%. Foi escolhido o que apresentou melhor RMSE de validação.

3.4 Abordagem de Score Balanceado para Seleção de Modelos

Para avaliar o equilíbrio entre desempenho e variância do modelo, fundamentado em (MICHELUCCI; VENTURINI, 2025), o indicador Composite Overfitting Score é definido por

$$\text{COS} = \alpha \frac{\text{Desempenho}_{\text{teste}}}{\text{Desempenho}_{\text{treino}}} + \beta \frac{\sigma_{\text{teste}}}{\sigma_{\text{treino}}}, \quad (3.3)$$

em que $\text{Desempenho}_{\text{treino}}$ e $\text{Desempenho}_{\text{teste}}$ denotam o RMSE, o MAE ou o R^2 para treino e teste, enquanto σ_{treino} e σ_{teste} denotam o desvio padrão das predições para treino e teste, e α e β ponderam a importância do desempenho médio (primeiro termo) e da variância do modelo (segundo termo). Neste trabalho, $\alpha = 0.5$ e $\beta = 0.5$ que dá igual importância para o desempenho e variância do modelo. Assim, $\text{COS} \approx 1$ indica um modelo ideal, no qual os erros de treino e teste coincidem, e a variância entre os *folds* durante o treinamento é consistente.

3.5 Interpretabilidade dos Modelos

Para além das métricas de desempenho, este trabalho também se concentra em extrair explicações para cada previsão dos modelos finais para os circuitos. Para a Regressão Linear Múltipla, a interpretabilidade foi obtida diretamente dos coeficientes do modelo. Como as variáveis estão na mesma escala, os coeficientes revelam o sentido, positivo ou negativo, e a intensidade da contribuição de cada uma ao tempo de volta previsto. Com isso, a MLR oferece uma interpretabilidade bem direta, demonstrando numericamente o que deixa a volta mais lenta ou mais rápida. Para o XGBoost, primeiro vem a importância de variáveis fornecida pelo próprio *booster*, que mede a melhoria média na função de perda proporcionada pelas divisões associadas a cada variável. Em seguida, os valores SHAP, calculados com o *TreeExplainer*, que fornece uma decomposição aditiva e exata da predição para modelos baseados em árvores. A partir dos valores SHAP, resumindo a importância e a direção do efeito de cada *feature*, foi proposta uma explicação diretamente local das previsões, por meio de um *force plot*, que detalha como cada variável desloca a predição em relação ao valor esperado, anotada com o contexto da volta (piloto, equipe, ano, número da volta e composto de pneu).

Para o modelo híbrido, o SHAP foi aplicado ao componente LSTM, responsável por prever o resíduo. As explicações do LSTM referem-se, portanto, a esse termo de correção, e não ao tempo de volta completo, cuja parcela dominante já é interpretada pela via do *baseline*. Como o LSTM opera sobre sequências temporais e a atribuição é feita sobre um modelo Keras⁶, foi utilizado o `GradientExplainer`, que estima os valores SHAP a partir dos gradientes da saída em relação ao tensor de entrada tridimensional ($\text{amostras} \times \text{passos temporais} \times \text{variáveis}$), tomando um subconjunto representativo de sequências como referência (*background*). As variáveis categóricas de alta cardinalidade (piloto e equipe) entram na rede por camadas de *embedding*, cujos canais são códigos inteiros não diferenciáveis. Consequentemente, o método não lhes atribui valor, e elas são reportadas explicitamente como não atribuíveis, ao passo que todas as variáveis numéricas e o composto de pneu (codificado por One-hot Encoding) recebem atribuição válida. Com os valores SHAP, foram produzidas os mesmos tipos de análises que foram produzidas para o XGBoost. De maneira adicional, foi proposta uma análise medidas das amostras para cada volta da sequência, quantificando o peso de cada uma das voltas na correção aprendida pela rede.

3.6 Considerações Finais

Este capítulo descreve a metodologia utilizada e aborda a preparação dos dados, a configuração experimental, os algoritmos utilizados, o critério de seleção de modelos e as ferramentas de interpretabilidade, com o foco de construir a base técnica da qual os resultados do Capítulo 4 são discutidos.

A Seção 3.1 é de suma importância, pois conecta dados concretos os fatores físicos e estratégicos que foram discutidos na fundamentação teórica do Capítulo 2. A extração restrita à corrida principal e ao recorte de 2022 a 2025 (Seção 3.1.1) traduz diretamente as justificativas apresentadas no capítulo anterior sobre a natureza da corrida e a estabilidade do regulamento técnico dentro de um mesmo ciclo. O pré-processamento garante que o alvo do modelo corresponda ao ritmo puro de corrida definido na Seção 2.2. A engenharia de *features* e a análise exploratória não apenas preparam variáveis para os modelos, mas também demonstram, através da matriz de correlação e da análise de *loadings* do PCA, as primeiras evidências de que cada circuito possui sua característica única, que será discutida diretamente ao responder a questão de pesquisa QP4.

A configuração experimental descrita na Seção 3.2 estabelece o protocolo de validação temporal que sustenta a resposta à QP1, ao adotar a janela expansiva com ordem cronológica preservada, o trabalho garante que nenhuma métrica reportada no Capítulo 4 tenha sido superestimada por vazamento de dados futuros, condição sem a qual a

⁶ <<https://keras.io/>>

comparação entre a abordagem agregada e a abordagem filtrada por piloto, central para a **QP1**, perderia validade.

A Seção 3.3 apresenta os três modelos comparados ao longo do trabalho. A Regressão Linear Múltipla e o XGBoost se estabelecem como os *baselines* e o modelo híbrido MLR-LSTM, cuja motivação e arquitetura em dois estágios foram detalhadas nessa seção, constitui a resposta direta à **QP2**, ao propor que o componente recorrente aprenda apenas o resíduo não capturado pelo componente linear, em vez de modelar o tempo de volta em sua totalidade. Dessa maneira, propõe-se uma tentativa de melhorar a acurácia dos modelos e contornar fraquezas de abordagens de aprendizado profundo.

A Seção 3.4 introduz o indicador COS como resposta metodológica direta à **QP3**. Sua necessidade decorre de uma limitação das métricas tradicionais discutida ao longo do capítulo. Isoladamente, elas descrevem o desempenho médio de um modelo, mas não capturam se esse desempenho é consistente entre o conjunto de treino e o de teste em termos de variância. Ao combinar desempenho e variância em um único indicador, o COS permite auxiliar na decisão entre modelos estatisticamente empatados no Capítulo 4, papel que nenhuma das métricas isoladas seria capaz de cumprir.

Por fim, a Seção 3.5 estabelece o protocolo de interpretabilidade que sustenta a **QP4**. Enquanto os coeficientes da MLR e os valores SHAP do XGBoost já constituem prática relativamente estabelecida na literatura revisada no Capítulo 2, a aplicação do **GradientExplainer** ao componente residual do LSTM é o que permite que o modelo híbrido não seja tratado como uma caixa-preta. Dessa maneira, é possível entender a contribuição da abordagem de aprendizado profundo na previsão, apresentando resultados explicáveis.

Em síntese, cada elemento apresentado neste capítulo, desde a filtragem de *outliers* à escolha do indicador COS, corresponde a uma decisão necessária para que a predição do tempo de volta e a extração de *insights* estratégicos, objetivos centrais deste trabalho, pudessem ser conduzidas de forma metodologicamente consistente. O capítulo seguinte apresenta os resultados obtidos a partir desse protocolo, retomando, na ordem em que foram propostas, as cinco questões de pesquisa que orientam este trabalho.

4 Resultados

Esse capítulo apresenta a descrição dos resultados deste trabalho. A Seção 4.1 se concentra em responder as três primeiras questões de pesquisas através das métricas dos resultados de cada um dos modelos, explorando principalmente RMSE, MAE, R^2 e COS. A Seção 4.2 foca em expandir as respostas encontradas na seção anterior através de uma análise temporal nos dados preditos para o conjunto de teste, definido aqui como *holdout*, apresentando estudos de casos de pilotos em determinadas corridas. A Seção 4.3 abre a discussão em respeito à **QP4**, extraindo através das ferramentas apresentadas anteriormente informações extras de cada modelo. A Seção 4.4 conecta a análise interpretativa com o contexto real, extraindo *insights* estratégicos para cada circuito. Por fim, a Seção 4.5 sumariza os resultados do trabalho e desenha a ponte para as considerações finais.

4.1 Comparação de Desempenho entre Modelos

Para tratar a **QP1**, a série temporal de tempos de volta foi organizada por circuito para cada piloto. Os dados são divididos em 80% para os conjuntos de treino e validação e 20% para o conjunto de teste. A Tabela 7 apresenta métricas de desempenho como RMSE, MAE e R^2 para MLR, XGBoost e LSTM. Além disso, essa tabela mostra o número de pilotos, o total de instâncias (i.e., m) e o número de instâncias por piloto. Em termos de RMSE e MAE, o LSTM apresenta os piores resultados para os circuitos do Bahrein, Itália e Hungria. O XGBoost apresenta o pior desempenho para Itália e Hungria. O efeito da limitação de dados é mais pronunciado no LSTM, por se tratar de uma rede neural profunda. Surpreendentemente, o XGBoost apresenta o melhor desempenho para Bahrein, Arábia Saudita e Estados Unidos. De forma geral, nenhum dos modelos atinge um bom ajuste, com $R^2 < 0$. Ainda assim, o número de instâncias por piloto é uma ordem de grandeza menor que m .

Nesse contexto, para avaliar o desempenho dos modelos com a solução proposta neste trabalho, os tempos de volta das temporadas para todos os pilotos são agregados formando uma única série temporal, o MLR, o XGBoost e o LSTM foram diretamente comparados nos cinco circuitos analisados. Os resultados de desempenho dos modelos nos conjuntos de treino e teste são apresentados na Tabela 8, na qual o tamanho ótimo da janela é apresentado entre parênteses antes do nome do modelo. Essa tabela apresenta as métricas de RMSE, MAE e R^2 para cada conjunto de dados indicado na coluna “Conjunto” (por exemplo, “Treino” ou “Teste”). As métricas de desempenho são reportadas com seus respectivos intervalos de confiança de 95% (I.C.). Assim, os modelos na Tabela 8 apresentam desempenho superior aos da Tabela 7, na qual a série temporal é filtrada por

Tabela 7 – Comparação do desempenho dos modelos filtrados por piloto e LSTM puro.

Circuito	Modelo	RMSE (95%C.I.)	MAE (95%C.I.)	R ² (95%C.I.)	#piloto	m	m/piloto
Bahrein	MLR	0.69 (0.48,0.90)	0.61 (0.41,0.82)	-1.40 (-3.46, 0.65)	29	3,161	109.00
	XGBoost	0.53 (0.43,0.63)	0.45 (0.36,0.54)	-0.11 (-0.56,0.34)			
	LSTM	1.26 (0.95,1.57)	1.15 (0.84,1.46)	-5.01 (-7.79,-2.23)			
Arábia Saudita Arabia	MLR	0.95 (0.43,1.46)	0.86 (0.35,1.36)	-13.95 (-34.98,7.09)	28	2,660	95.00
	XGBoost	0.69 (0.55,0.83)	0.60 (0.47,0.72)	-3.45 (-6.75,-0.16)			
	LSTM	0.85 (0.68,1.02)	0.75 (0.60,0.90)	-6.28 (-10.76,-1.80)			
EUA	MLR	2.04 (0.74,3.34)	1.92 (0.63, 3.22)	-156.48 (-365.34,52.38)	29	3,063	105.60
	XGBoost	0.64 (0.47,0.81)	0.54 (0.38,0.69)	-1.13 (-2.15,-0.12)			
	LSTM	1.41 (1.17,1.65)	1.31 (1.08,1.55)	-10.79 (-16.55,-5.03)			
Itália	MLR	0.65 (0.45,0.84)	0.57 (0.39,0.75)	-4.97 (-9.89,-0.05)	30	2,994	99.80
	XGBoost	0.80 (0.63,0.96)	0.69 (0.54,0.84)	-4.38 (-7.13,-1.64)			
	LSTM	1.96 (1.58,2.34)	1.91 (1.53,2.29)	-36.81 (-49.24,-24.39)			
Hungria	MLR	0.63 (0.44,0.81)	0.54 (0.36,0.72)	-0.10 (-0.72,0.53)	29	4,213	145.30
	XGBoost	0.84 (0.68,1.01)	0.70 (0.55,0.85)	-0.70 (-1.50,0.09)			
	LSTM	1.55 (1.32,1.77)	1.41 (1.20,1.63)	-5.60 (-8.05,-3.15)			

piloto, reduzindo o número de instâncias. Por exemplo, na Tabela 7, para o circuito dos EUA, os modelos MLR e LSTM obtiveram um MAE de 1.92 segundos e 1.31 segundos, respectivamente. Já na Tabela 8, para o mesmo circuito, o MLR obtém um MAE de 0.26 segundos e o MLR-LSTM obtém um MAE de 0.27 segundos, cujo intervalo de confiança difere significativamente dos resultados apresentados na Tabela 7.

Ao tratar a **QP2**, foi proposta uma estratégia para melhorar a acurácia da predição do tempo de volta por meio do modelo híbrido MLR-LSTM. A Tabela 8 destaca em vermelho todos os circuitos nos quais nossa solução atinge o melhor desempenho, com a menor média em todos os resultados de treino. Esse cenário se altera quando se analisa os resultados dos conjuntos de teste, nos quais o melhor desempenho é destacado em verde. No entanto, há um empate entre todos os modelos, pois não há resultados significativamente diferentes (i.e., os I.C. se sobrepõem). A exceção é o circuito da Itália, na qual o modelo híbrido apresenta o melhor desempenho, com resultados para todas as métricas significativamente diferentes dos demais modelos (seus I.C. estão em vermelho). Ao observar o RMSE e o MAE, o MLR-LSTM é pelo menos 11% e 27% mais preciso em relação ao MLR e ao XGBoost, respectivamente. O empate observado nos demais circuitos pode estar relacionado à capacidade de generalização do modelo no conjunto de teste, que consiste em registros de 2025. Uma inspeção manual revela que essas distribuições de dados diferem daquelas presentes no conjunto de treino.

Para responder à **QP3**, também é reportado, na Tabela 8, o desvio padrão dos resíduos para os conjuntos de treino e teste, a fim de ilustrar a variabilidade do modelo; σ_{teste} e σ_{treino} são utilizados para calcular o COS. A Tabela 8 reporta o COS_{RMSE} , o COS_{MAE} e o COS_{R^2} , respectivamente. Esses escores são utilizados para equilibrar o desempenho médio com a variabilidade do modelo, auxiliando na identificação do modelo mais adequado para cada circuito. Nessa tabela, os menores escores de COS são destacados em azul. De acordo com o COS, o melhor modelo para prever os tempos de volta no circuito da Itália é o híbrido MLR-LSTM, com os menores escores de COS. Além disso, apenas seu RMSE e MAE diferem significativamente dos demais modelos. No circuito dos EUA, o

Tabela 8 – Comparação do desempenho entre os modelos.

Circuito	Conjunto	Modelo	σ	RMSE (95%C.I.)	COS _{RMSE}	MAE (95%C.I.)	COS _{MAE}	R ² (95%C.I.)	COS _{R²}
Bahrein	Treino	MLR (EW 5%)	0.3059	0.36 (0.25,0.47)	–	0.30 (0.20,0.39)	–	0.64 (0.37,0.91)	–
		XGBoost (EW 30%)	0.2821	0.29 (0.24,0.33)	–	0.22 (0.19,0.26)	–	0.88 (0.86,0.91)	–
		Hybrid MLR-LSTM	0.3168	0.25 (0.23,0.27)	–	0.19 (0.18,0.21)	–	0.91 (0.90,0.93)	–
	Teste	MLR	0.3140	0.31 (0.30,0.33)	0.95	0.25 (0.23,0.26)	0.93	0.88 (0.86,0.90)	0.88
		XGBoost (EW 30%)	0.3122	0.34 (0.32,0.36)	1.15	0.27 (0.26,0.29)	1.16	0.86 (0.83,0.88)	1.07
		Hybrid MLR-LSTM	0.3164	0.32 (0.30,0.34)	1.25	0.24 (0.23,0.26)	1.26	0.87 (0.85,0.89)	1.15
Arábia Saudita	Treino	MLR (EW 10%)	0.3406	0.37 (0.34,0.39)	–	0.29 (0.27,0.32)	–	0.80 (0.76,0.84)	–
		XGBoost (EW 50%)	0.3168	0.32 (0.27,0.36)	–	0.24 (0.20,0.29)	–	0.87 (0.81,0.94)	–
		Hybrid MLR-LSTM	0.3088	0.30 (0.28,0.33)	–	0.24 (0.22,0.26)	–	0.94 (0.93,0.95)	–
	Teste	MLR	0.3185	0.34 (0.31,0.36)	0.92	0.26 (0.24,0.28)	0.91	0.91 (0.90,0.93)	0.91
		XGBoost (EW 50%)	0.3175	0.33 (0.30,0.35)	1.02	0.25 (0.23,0.27)	1.01	0.92 (0.90,0.93)	0.98
		Hybrid MLR-LSTM	0.3186	0.33 (0.31,0.35)	1.06	0.25 (0.24,0.27)	1.05	0.92 (0.90,0.93)	1.03
EUA	Treino	MLR (EW 45%)	0.3518	0.36 (0.34,0.39)	–	0.28 (0.26,0.31)	–	0.83 (0.80,0.86)	–
		XGBoost (EW 5%)	0.3117	0.32 (0.30,0.33)	–	0.26 (0.24,0.27)	–	0.83 (0.81,0.86)	–
		Hybrid MLR-LSTM	0.3186	0.32 (0.30,0.34)	–	0.25 (0.23,0.27)	–	0.91 (0.89,0.92)	–
	Teste	MLR	0.3330	0.33 (0.32,0.35)	0.93	0.26 (0.25,0.28)	0.93	0.87 (0.85,0.88)	0.95
		XGBoost (EW 5%)	0.3378	0.36 (0.34,0.38)	1.11	0.28 (0.27,0.30)	1.10	0.84 (0.82,0.86)	1.04
		Hybrid MLR-LSTM	0.3301	0.34 (0.32,0.36)	1.05	0.27 (0.25,0.28)	1.05	0.86 (0.84,0.88)	1.04
Itália	Treino	MLR (EW 5%)	0.2930	0.36 (0.30,0.41)	–	0.29 (0.24,0.35)	–	0.61 (0.42,0.79)	–
		XGBoost (EW 50%)	0.3309	0.36 (0.20,0.51)	–	0.28 (0.16,0.40)	–	0.80 (0.72,0.88)	–
		Hybrid MLR-LSTM	0.2675	0.27 (0.25,0.29)	–	0.21 (0.19,0.22)	–	0.90 (0.88,0.92)	–
	Teste	MLR	0.2715	0.33 (0.31,0.35)	0.92	0.27 (0.25,0.28)	0.92	0.84 (0.82,0.87)	0.82
		XGBoost (EW 50%)	0.3413	0.43 (0.40,0.46)	1.12	0.33 (0.30,0.35)	1.10	0.74 (0.71,0.76)	1.06
		Hybrid MLR-LSTM	0.2657	0.29 (0.28,0.31)	1.05	0.24 (0.22,0.25)	1.07	0.87 (0.85,0.89)	1.01
Hungria	Treino	MLR (EW 45%)	0.3238	0.33 (0.30,0.36)	–	0.26 (0.23,0.28)	–	0.88 (0.84,0.91)	–
		XGBoost (EW 40%)	0.3201	0.33 (0.30,0.37)	–	0.26 (0.23,0.29)	–	0.85 (0.81,0.90)	–
		Hybrid MLR-LSTM	0.3110	0.31 (0.29,0.33)	–	0.23 (0.22,0.25)	–	0.90 (0.89,0.92)	–
	Teste	MLR	0.3412	0.34 (0.33,0.36)	1.05	0.27 (0.25,0.28)	1.05	0.91 (0.90,0.92)	1.01
		XGBoost (EW 40%)	0.5466	0.59 (0.56,0.62)	1.74	0.44 (0.41,0.46)	1.71	0.73 (0.71,0.75)	1.44
		Hybrid MLR-LSTM	0.3462	0.35 (0.33,0.37)	1.12	0.27 (0.26,0.29)	1.14	0.90 (0.89,0.91)	1.06

modelo híbrido também apresenta escores de COS menores, em torno de 1.05 em todas as métricas de desempenho. Em contrapartida, o pior COS do MLR-LSTM ocorreu nos circuitos do Bahrein e da Hungria, nos quais o modelo apresentou um equilíbrio ruim entre desempenho médio e variabilidade. Nesses circuitos com empates estatísticos, utilizamos o COS como critério de desempate, de modo que o MLR obtém o melhor COS. Na Arábia Saudita, o XGBoost alcança o melhor COS, seguido de perto pelo MLR-LSTM, com valores comparáveis. Em contrapartida, o menor escore de COS para o XGBoost foi observado na Hungria, com um equilíbrio muito ruim.

Foi criado o gráfico de Função de Densidade Acumulada, do inglês, Cumulative Distribution Function (CDF), para as métricas MAE, RMSE e R² do desempenho dos modelos separados por cada piloto. No caso da Arábia Saudita, na Figura 17, foi observado que o XGBoost apresenta um desempenho superior em alguns casos, mas a vantagem não se sustenta ao longo de toda a distribuição, com as curvas se cruzando na cauda. Nos Estados Unidos (Figura 18), o cenário é distinto, pois o baseline da Regressão Linear Múltipla obtém os menores erros para a maioria dos pilotos, enquanto o modelo híbrido acompanha de perto e o XGBoost fica atrás. No entanto, nenhum modelo se destaca de forma consistente. Para a Itália, no entanto, o modelo híbrido se sobressai de maneira mais evidente, com erros menores e R² superior ao longo de toda a CDF.

A Figura 20 representa visualmente os resultados de teste da Tabela 8, permitindo comparar os três modelos e seus intervalos de confiança de 95% em cada circuito. Nos circuitos do Bahrein, da Arábia Saudita e dos Estados Unidos, as barras dos três modelos

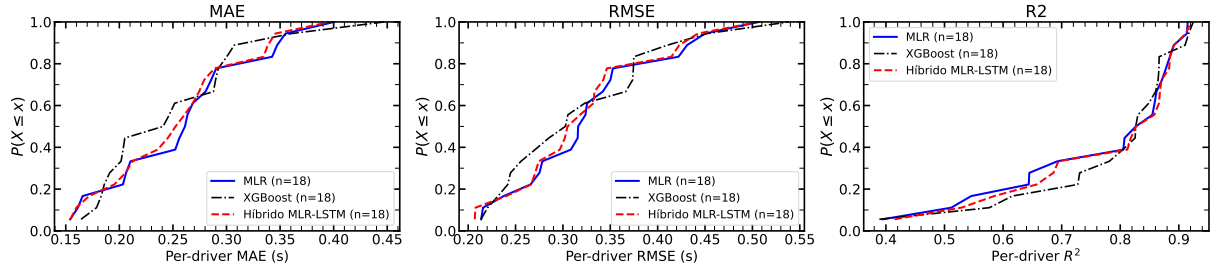


Figura 17 – CDF das métricas por piloto para o Grande Prêmio da Arábia Saudita.

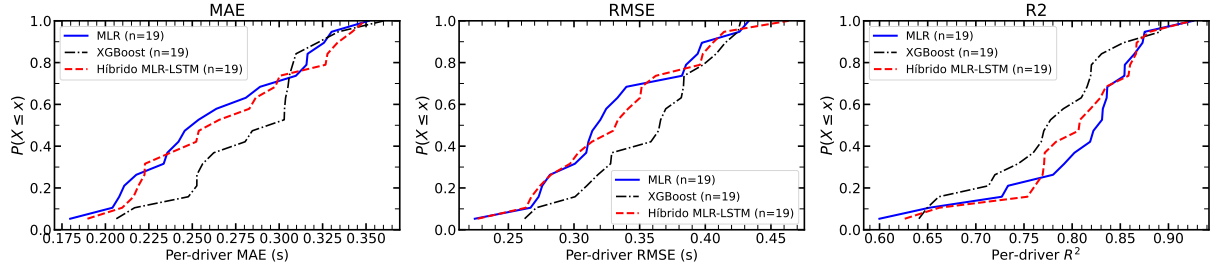


Figura 18 – CDF das métricas por piloto para o Grande Prêmio dos Estados Unidos.

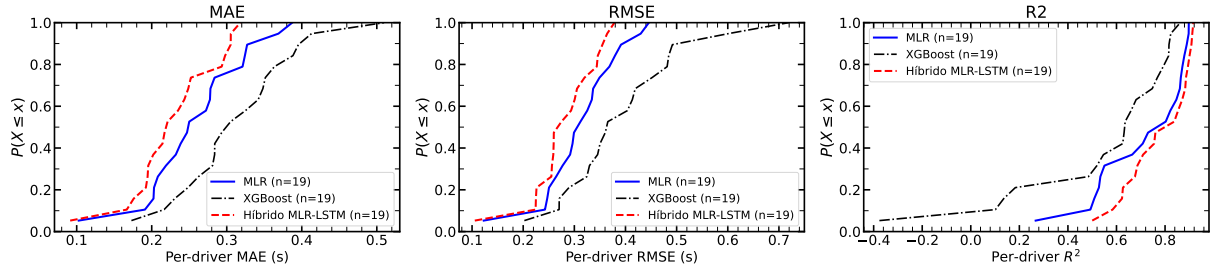


Figura 19 – CDF das métricas por piloto para o Grande Prêmio da Itália.

são praticamente indistinguíveis e seus intervalos de confiança se sobrepõem em todas as métricas, caracterizando o empate estatístico já apresentado anteriormente. Na Itália porém, o modelo híbrido MLR-LSTM apresenta intervalos que não se sobrepõem aos dos *baselines*, confirmando-o como o único circuito em que a estratégia híbrida se destaca de maneira estatisticamente significativa. Essa figura também evidencia a fragilidade do XGBoost fora dos circuitos em que ele se ajusta bem: na Hungria e na Itália seu erro cresce de forma acentuada (na Hungria, o MAE chega a 0.44 seg. e o RMSE a 0.59), enquanto o MLR e o híbrido permanecem estáveis em torno de 0.25–0.35 segundos. Esse comportamento reforça que, sob escassez de dados e deslocamento de distribuição entre temporadas, o baseline linear é quem mais se destaca. O Híbrido consegue agregar quando a natureza temporal do problema permite, como em Monza na Itália.

A Seção C.4 do Apêndice C demonstra, por meio das tabelas de desvio padrão, a diferença entre treino e teste para cada um dos modelos, evidenciando a consistência e melhora apresentada pelo modelo híbrido na Itália. Também é possível ver a queda que acontece para o Grande Prêmio da Bahrein no *holdout*.

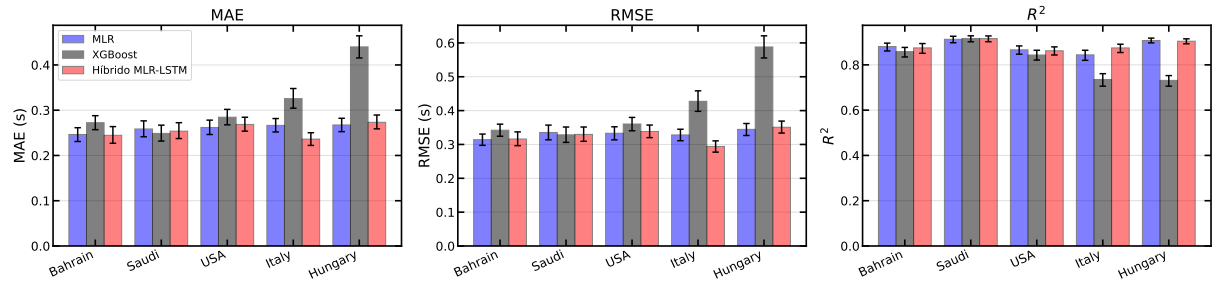


Figura 20 – Desempenho agregado no conjunto de teste (*holdout* de 2025) por circuito e modelo.

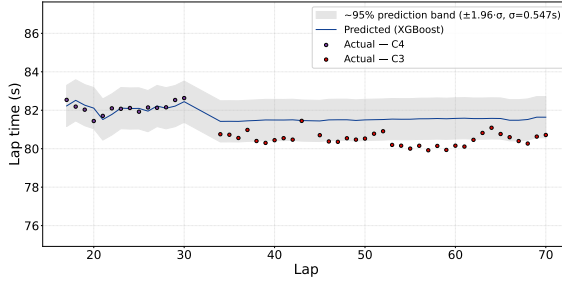
4.2 Análise Temporal do Holdout

O *holdout* tem como propósito principal simular um teste real do que seria uma previsão aplicada em uma corrida de Fórmula 1. Além dos contrapontos já mencionados na **QP1** sobre a limitação de dados ao separar por série temporal de piloto e circuito, essa abordagem limita ainda mais os dados, mas traz a análise mais valiosa. Mais do que as métricas agregadas, interessa aqui o comportamento volta a volta das séries preditas, que responde na prática à **QP2** e à **QP3**. Considerando que o *holdout* reserva a maior parte das voltas das corridas de 2025 para cada uma das pistas, essa análise se aproxima da Fórmula 1 real, onde equipes utilizam dados de anos anteriores e do contexto inicial daquela respectiva corrida para prever o restante.

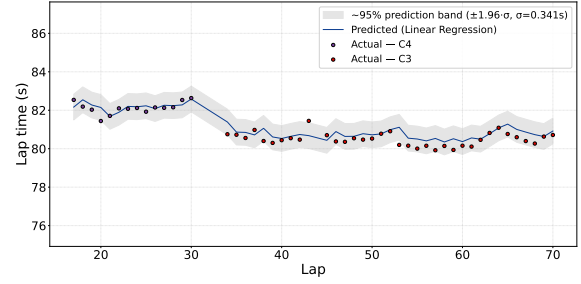
As Figuras 21, 22 e 23 apresentam séries temporais de tempo de volta no *holdout* organizadas em três contexto pares, cada um contrastando os três modelos sobre a mesma corrida: Norris na Hungria, Verstappen na Itália e Russell nos Estados Unidos. Em cada gráfico, os pontos coloridos indicam os tempos reais por composto Pirelli, a linha é a predição e a região sombreada é a banda de predição de aproximadamente 95%.

No primeiro contexto, a corrida de Norris na Hungria ilustra o efeito de uma mudança de estratégia. Em 2025 o piloto inglês apostou em uma parada única, trocando do C4 para o C3 em torno da volta 30, o que difere do padrão de duas paradas dos anos anteriores, seguido, por exemplo, por seu companheiro Piastri (C3 na volta 22 e nova parada próximo à volta 50, usufruindo de pneus mais novos e rápidos). Essa estratégia rendeu a vitória a Norris, ainda que com tempos de volta mais lentos que os de seu perseguidor. Como o *stint* longo em C3 não corresponde ao histórico de treino, o XGBoost (Figura 21(a)) falha nesse trecho: projeta tempos praticamente constantes em torno de 81.5 seg., enquanto Norris rodava cerca de 80 seg., deixando os pontos reais sistematicamente abaixo da banda, que é larga ($\sigma = 0.547$). A Regressão Linear (Figura 21(b)), por outro lado, acompanha a queda de ritmo no início do C3 e mantém os tempos reais majoritariamente dentro de uma banda mais estreita ($\sigma = 0.341$), o que é coerente com seu desempenho superior ao do XGBoost nesse circuito. No cenário das voltas do Norris no *holdout*, a abordagem híbrida, visualmente demonstrada pela Figura 21(c), não apresenta ganho

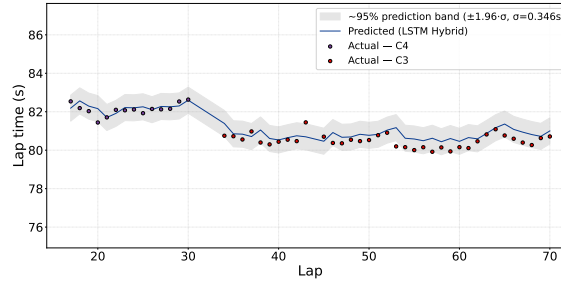
significativo. Pelo contrário, as voltas tem erros um pouco maiores do que seu *baseline*.



(a) Norris - Hungria (XGBoost).



(b) Norris - Hungria (Regressão Linear).

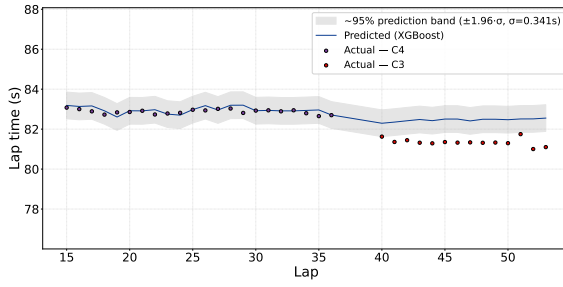


(c) Norris - Hungria (Modelo Híbrido).

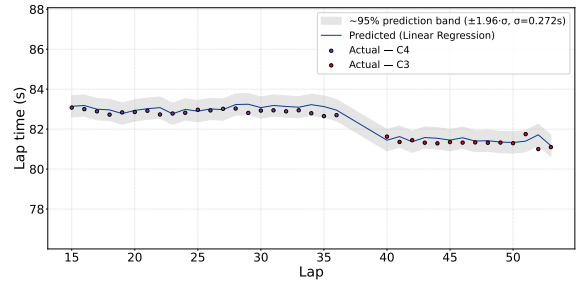
Figura 21 – Séries temporais comparativas para o *holdout* na Hungria para o piloto Norris.

Por outro lado, na segunda comparação, a corrida de Verstappen na Itália evidencia o ganho do modelo híbrido. Sob o XGBoost (Figura 22(a)), a predição acompanha bem o primeiro *stint* em C4, mas erra o *stint* final em C3. O modelo projeta tempos em torno de 82.4 seg., enquanto Verstappen imprime um ritmo próximo de 81.3 seg., deixando os pontos reais abaixo da predição. Esse desempenho individual forte e inesperado para o ano é justamente o que os modelos *baseline* não conseguem captar. O modelo híbrido (Figura 22(c)) captura o ritmo de corrida forte da volta 40 em diante e passa a seguir os tempos reais do *stint* final, com uma banda mais estreita ($\sigma = 0.266$) contra 0.341 seg. do XGBoost). Esse contraste, no nível de uma única corrida, é a contrapartida qualitativa da vantagem estatística do híbrido na Itália observada na Tabela 8. A Figura 22(b) apresenta a série temporal para o modelo MLR. Embora a forma da série seja semelhante, o modelo híbrido apresenta erros ligeiramente menores, como evidenciado pela banda de predição ($\sigma = 0.272$ versus $\sigma = 0.266$).

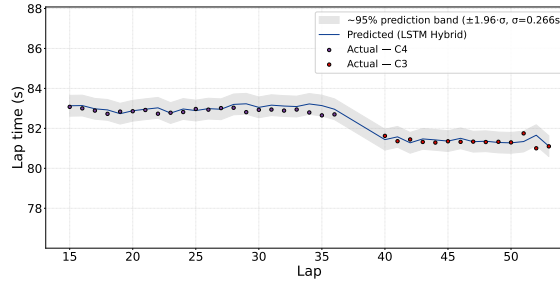
Por fim, a terceira e última comparação proposta diz respeito à corrida de George Russell no GP dos Estados Unidos. Aqui, existe um cenário de maior equilíbrio entre os modelos. O piloto da Mercedes teve um desempenho extremamente consistente durante os anos de 2022 a 2024, normalmente finalizando a corrida na quinta e sexta colocações, respectivamente. Dessa maneira, é sensato imaginar que os três modelos consigam compreender a natureza do comportamento daquele piloto naquela corrida em específico. A Figura referente à série temporal do XGBoost 23(a) tem uma natureza semelhante às outras duas, referentes à, respectivamente, MLR (Figura 23(b)) e o híbrido (Figura 23(c)).



(a) Verstappen - Itália (XGBoost).



(b) Verstappen - Itália (Regressão Linear).



(c) Verstappen - Itália (Modelo Híbrido).

Figura 22 – Séries temporais comparativas para o *holdout* na Itália para o piloto Verstappen.

Mesmo que a Tabela 8 aponte para uma superioridade em termos de média para a MLR no teste, no cenário isolado de Russell, a abordagem híbrida se mostra superior, apresentando uma borda mais estreita ($\sigma = 0.330$) que o *baseline*. Dessa maneira, a análise temporal do *holdout* indica diretamente para uma coisa: mesmo que o desempenho geral de um modelo não seja o melhor, ele pode ainda sim superar os outros em contextos de corrida específicos. O modelo híbrido consegue prever bem para um piloto consistente na pista dos Estados Unidos, mas outros casos podem prejudicar o desempenho geral.

Para uma contextualização mais detalhada, os resultados de cada uma das corridas, inclusive as corridas mencionadas para a análise da série temporal, podem ser observados no Apêndice B.

Ainda para o *holdout*, é necessário caracterizar a diferença da natureza das voltas que são utilizadas para treino, validação e teste. O gráfico de densidade representado pela Figura 24 ajuda a entender a diferença entre cada um desses cenários, se baseando no que o modelo híbrido utiliza para seu treino e teste. Para o Grande Prêmio da Itália, existe uma evolução temporal bem marcada: o *holdout*, que é exclusivo da corrida de 2025, tem as voltas mais rápidas, com a maioria girando na casa dos 82 segundos. Logo depois, a validação do modelo, que pega parte da corrida de 2025 e o final da corrida de 2025 e depois o restante, que seriam as voltas de treino nos anos anteriores. Para alguns circuitos, como o Bahrein, há uma confusão temporal não esperada: os resultados da validação para a pista são ótimos e pode-se observar através do gráfico de densidade que há a evolução dos tempos de volta que são reservados para o treino e validação. No entanto, quando se nota

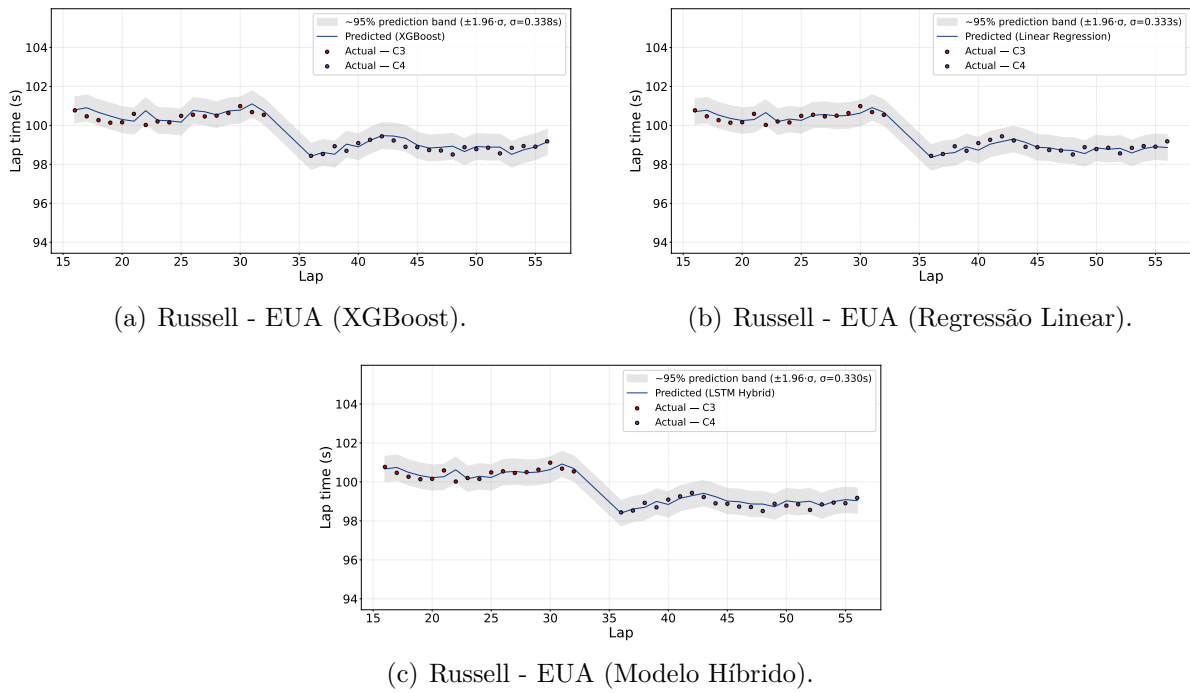


Figura 23 – Séries temporais comparativas para o *holdout* os EUA para o piloto Russell.

a densidade das voltas do *holdout*, percebe-se que elas são mais lentas que o imaginado, evidenciando que o ano de 2024 teve voltas mais rápidas que 2025, o que interfere no desempenho do modelo. Isso pode ser validado na Seção A.3 do Apêndice A, através da Tabela 20, que mostra que os tempos dos médios dos pilotos ficaram mais lentos.

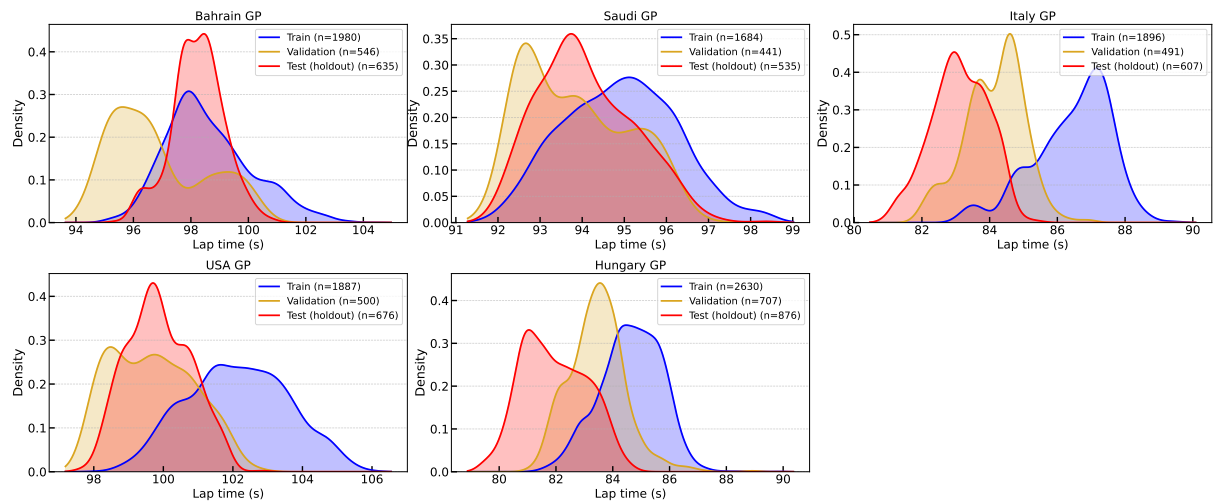


Figura 24 – Densidade do tempo de volta em cada um dos circuitos.

4.3 Análise Interpretativa dos Modelos

Com o objetivo de avaliar as individualidades de cada circuito, foi conduzida uma análise mais abrangente, a partir da análise exploratória de dados até a comparação entre

os resultados obtidos pelos três tipos de modelos utilizados. A Matriz de Correlação e a análise do PCA Loadings foram aplicados para identificar relações lineares prévias e o agrupamento de desempenho das equipes e pilotos. Essa abordagem prepara o terreno para identificar as nuances aerodinâmicas, climáticas e mecânicas específicas de cada traçado estudado.

Primeiramente, a Regressão Linear Múltipla revela os *insights* mais fáceis de entender. A Figura 25 apresenta os coeficientes da MLR para as pistas da Hungria (25(a)) e dos Estados Unidos (25(b)). Em ambos os casos, assim como nas outras pistas, a variável de maior coeficiente é o tempo de volta anterior. Ela se destaca como a variável mais consistente para explicar o tempo de volta atual, comprovando a forte dependência temporal no comportamento das voltas, o que valida a abordagem proposta neste trabalho que respeita a sequência das voltas, tanto para as abordagens de janela expansiva, quanto para a escolha do LSTM. No entanto, é possível observar resultados diferentes. A vida do pneu (TyreLife) aparece em segundo lugar dos coeficientes positivos para a Hungria, enquanto nos Estados Unidos a Pressão Atmosférica é quem ocupa esse lugar. Para variáveis de coeficientes negativos, ou seja, que abaixam o tempo de volta, o número da volta aparece presente fortemente em todas as pistas, normalmente como o segundo maior coeficiente absoluto ou próximo disso. Algumas equipes e pilotos alternam sua presença nesses coeficientes, tal como sua importância. Como observado para a pressão atmosféricas em Austin (25(b)), cada circuito tem determinada importância diferente para variáveis climáticas. Para a própria pista do Texas, a Temperatura da Pista é a segunda variável de maior coeficiente, enquanto para a Hungria (25(a)) essa importância é menor, mas aparece. Os resultados iniciais da interpretabilidade do modelo baseline corroboram para a análise dos outros modelos, começando a responder a QP4.

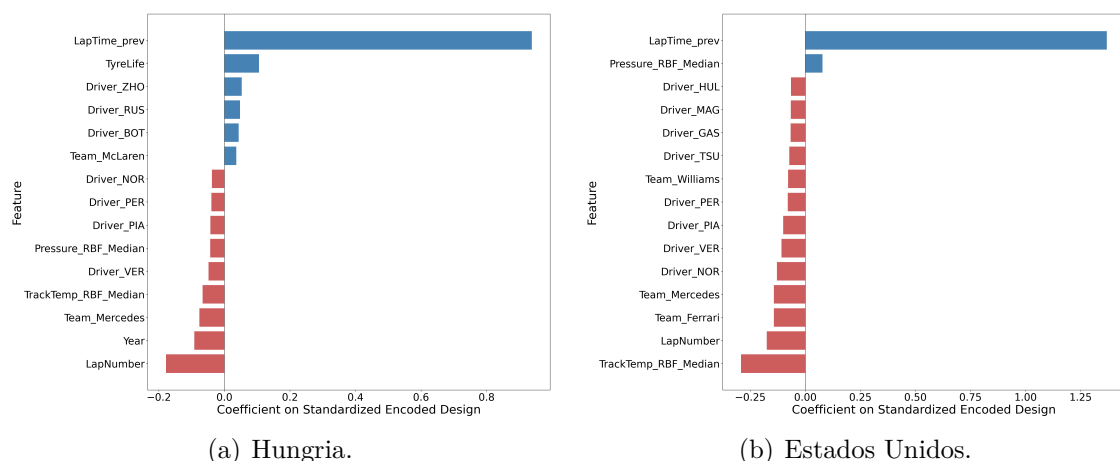


Figura 25 – Coeficientes da Regressão Linear Múltipla.

O modelo XGBoost oferece uma visão adicional aos baselines, com a análise do ganho de importância das variáveis e dos valores SHAP. O ganho trás uma análise direta semelhante aos coeficientes da MLR. Na Figura 26 são apresentados os ganhos para as

pistas da Arábia Saudita (26(a)) e da Itália (26(b)). Através deles, é possível ver, mais uma vez, o tempo de volta anterior como a variável que mais se destaca. A variável só não é a mais importante para o Grande Prêmio da Itália, onde o ano se destaca com uma importância ainda maior. As duas pistas apresentadas apresentam a maior distribuição de importância; nas outras três, a importância se concentra diretamente na variável autorregressiva. Nesse contexto, ele é melhor balanceado, por exemplo, na Itália, com Ano e o Tempo Anterior dividindo sua importância, e na Arábia Saudita, com o tempo anterior, ano, a equipe Red Bull Racing e o número da volta, todos mais balanceados.

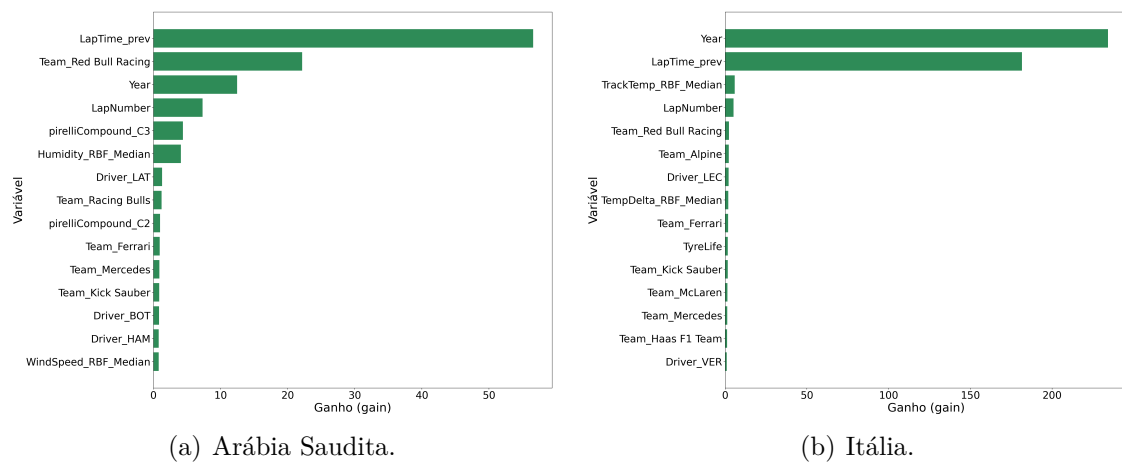


Figura 26 – Ganho de importância no XGBoost.

A análise do ganho sozinha consegue explicar as diferenças entre os resultados obtidos na MLR, mas a análise de valores SHAP expande bem mais essa discussão. A Figura 27 apresenta os SHAP Values para a validação e teste do modelo XGBoost para o Grande Prêmio da Arábia Saudita, o modelo que apresentou melhor desempenho entre os do tipo *ensemble*. Na representação, pode ser vista de maneira muito mais clara o impacto de cada variável na predição do modelo. As voltas de validação se distribuem ao longo do anos através das janelas expansivas, então o cenário visto na Figura 27(a) representa mais um panorama geral, enquanto a Figura 27(b) diz respeito ao teste final, ou seja, o *holdout*, que conta exclusivamente com voltas da corrida de 2025. É interessante destacar a força da Red Bull Racing, destacada anteriormente na importância para essa pista. Principalmente nos anos entre 2022-2024, a equipe foi extremamente dominante na pista, sendo sinônimo de velocidade. Variáveis climáticas, como umidade do ar, velocidade do vento e o delta de temperatura também tem impacto consistentemente forte na saída do modelo, seja ele positivo ou negativo. Os gráficos do valores SHAP servem para, em todos os circuitos, expandir melhor as informações de importância, informando, além do ganho, qual impacto real aquelas variáveis tem na predição. De maneira adicional, o *force plot* foi utilizado para decompor localmente a colaboração de cada variável em uma volta específica, como apresentado pela Figura 28.

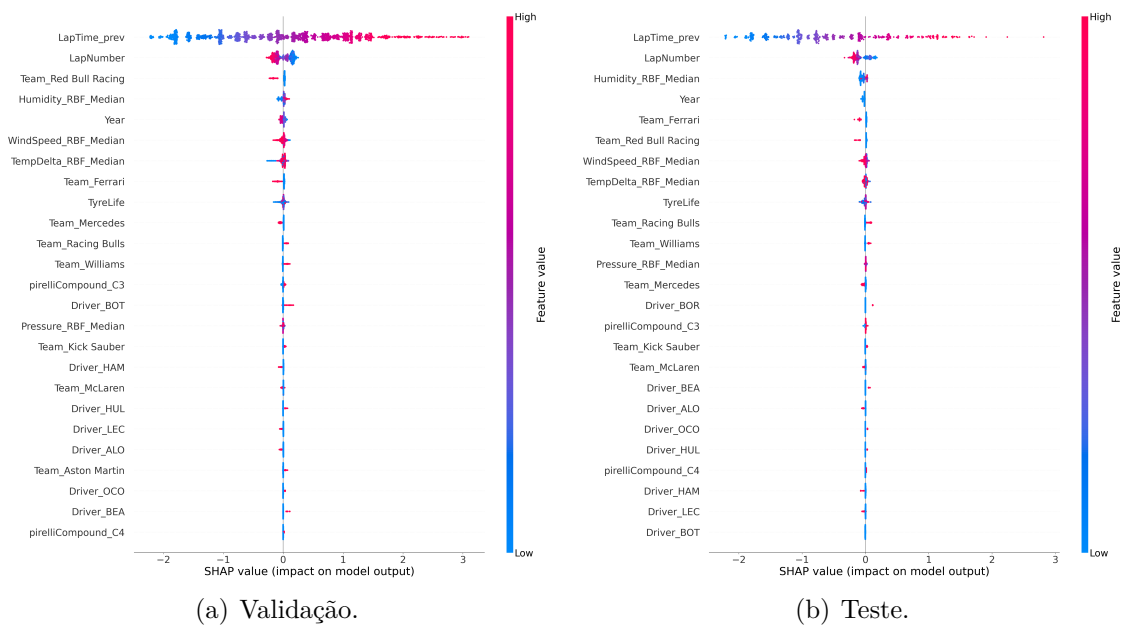


Figura 27 – SHAP *Values* para o Grande Prêmio da Arábia Saudita.

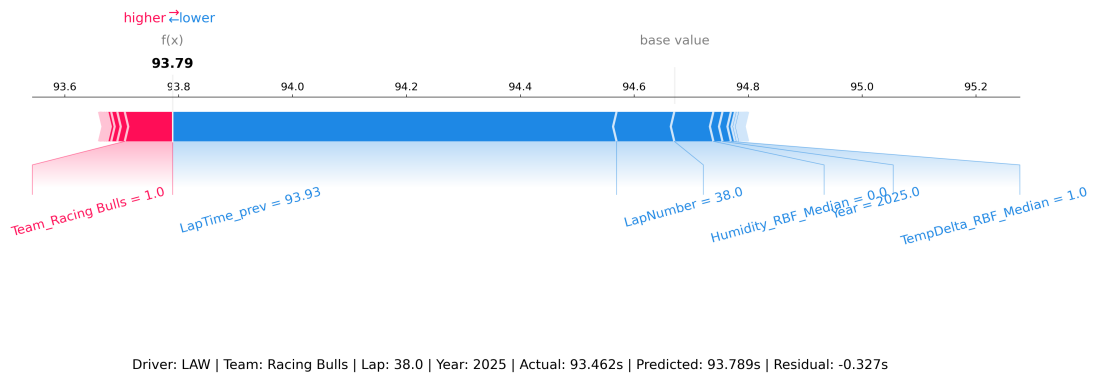


Figura 28 – Decomposição SHAP (*force plot*) da predição de uma volta no GP da Arábia Saudita no *holdout*.

Por fim, o modelo híbrido possibilita a análise adicional da interpretabilidade da predição do resíduo. Se, no começo desse capítulo, os coeficientes da Regressão Linear Múltipla foram destrinchados, o que esse resíduo capturado através do LSTM ajuda a explicar, por exemplo, para o caso da pista da Itália? Através da análise de valores SHAP, estendida para o LSTM, é possível identificar o que essa correção traz. A Figura 29 apresenta os valores SHAP para o Grande Prêmio da Itália na validação e no *holdout*. Nos coeficientes da MLR para essa pista, o composto do pneu é praticamente ignorado: as variáveis `pirelliCompound_C3`, `C4` e `C5` ocupam a 45^a, a 46^a e a 36^a posições entre as 50 do modelo, com coeficientes absolutos inferiores a 0.008, e a velocidade do vento é a penúltima colocada. No resíduo capturado pelo LSTM, contudo, essas mesmas variáveis lideram o *ranking* do *holdout* (Figura 29(b)): `pirelliCompound_C4`, `TyreLife` e a velocidade do vento (0.0149) são os três canais mais influentes. O híbrido, portanto, não apenas reduz o erro: ele o reduz recuperando um efeito de degradação e de composto que a formulação

linear, com um único coeficiente por variável, é incapaz de representar. É relevante destacar também o contraste entre os dois blocos. Na validação, o resíduo é dominado por variáveis climáticas, ao passo que, no *holdout* de 2025, o composto e a vida do pneu assumem a liderança, com magnitudes cerca de duas vezes maiores. Esse deslocamento sugere que a componente não linear ganha importância exatamente no bloco em que a baseline extrapola para uma temporada não vista, o que é coerente com o ganho de desempenho do híbrido sobre a MLR nessa pista. O perfil de recência é descrito pela Figura 31 reforça o caráter local dessa correção para a Itália (31(a)): a volta alvo concentra 48.3% da magnitude média dos valores SHAP, contra 28.4% em $t - 1$ e 23.3% em $t - 2$, indicando que o resíduo é explicado sobretudo pelo estado imediato do carro, e não por um histórico longo, o que também justifica o comprimento de sequência curto (três voltas) selecionado pela otimização para a Itália. Para o Bahrein, a sequência longa encontrada (17 voltas) pela busca de hiperparâmetros não é tão aproveitada. A Figura 31(b) demonstra a magnitude dos SHAP *Values*, que não é significativa para voltas iniciais da sequência. No entanto, mesmo com o gráfico ajudando a entender a diferença dos dois problemas em termos de memória curta ou longa, o resíduo é quase sempre explicado pelas condições encontradas na volta alvo. O *force plot* da Figura 30 decompõe essa lógica em uma volta específica do *holdout*.

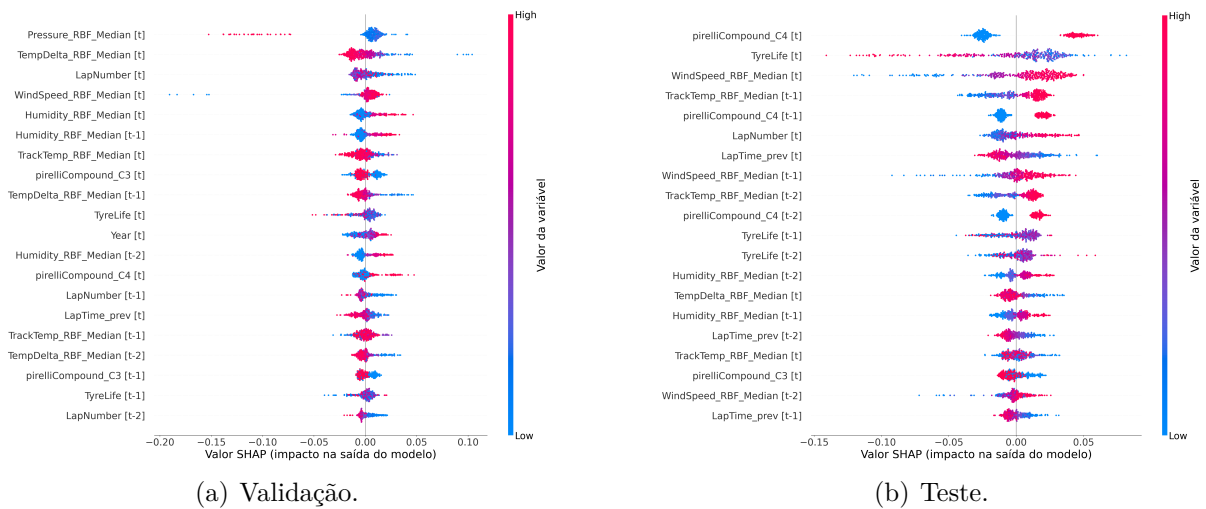


Figura 29 – SHAP *Values* para o Grande Prêmio da Itália no modelo LSTM.

As informações obtidas através dessas extrações demonstram a importância da sumarização de cada uma dessas informações na seção posterior, o que ajudará a identificar *insights* e a responder à **QP4** de maneira sucinta. De maneira geral, um contexto maior para a discussão levantada aqui aparece no Apêndice C, apresentado números completos para análise de coeficientes, importância das variáveis e SHAP *Values*.

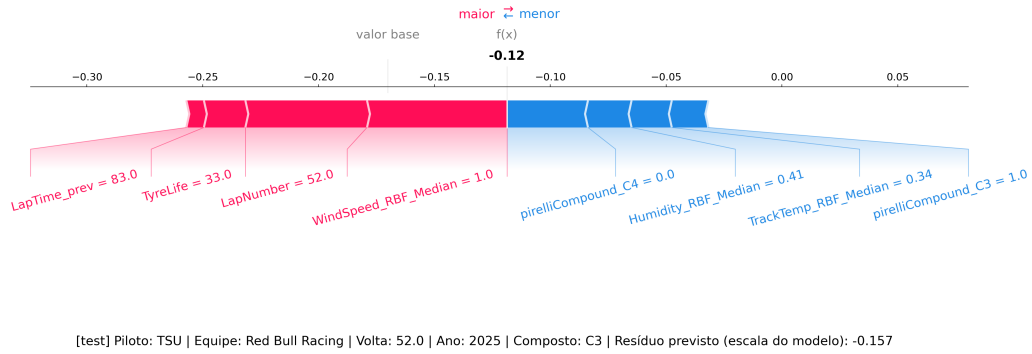


Figura 30 – Decomposição SHAP (*force plot*) da previsão de uma volta no GP da Itália no *holdout*.

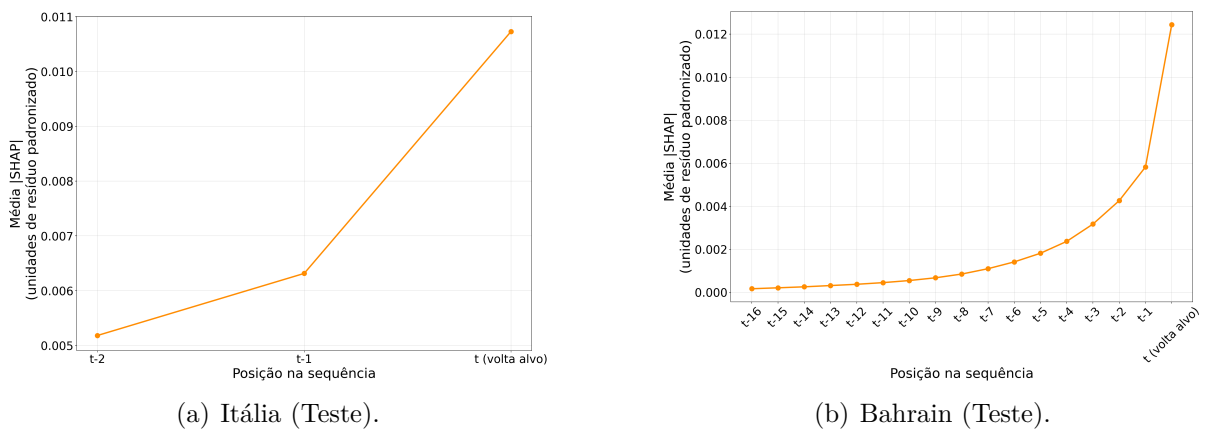


Figura 31 – Perfil de recência para o modelo LSTM.

4.4 Insights Estratégicos por Circuito

A seção anterior se concentra em responder quais *features* mais impactam cada um dos tipos de modelos nesses circuitos. No entanto, a interpretação de cada uma dessas informações é importante. Só é possível extrair *insights* estratégicos que auxiliem na tomada de decisão para as corridas e responder completamente a **QP4** se o trabalho se propor a separar cada uma das explicações em respostas coesa, que expliquem detalhes climáticos, mecânicos e temporais para cada um dos cenários.

A Tabela 9 consolida, por circuito, aquilo que os três modelos apontam. O padrão comum a todas as pistas é a dominância da componente autorregressiva: o tempo da volta anterior é o maior coeficiente da MLR e o maior ganho do XGBoost em quatro dos cinco circuitos. No entanto, a dominância dessa *feature* não nos impede de capturar a assinatura de cada um dos circuitos. A Hungria se destaca pela sensibilidade à pressão atmosférica e pela alta degradação, de todas as pistas escolhidas, é aquela que tem a menor velocidade média e necessita de grande carga aerodinâmica. Monza é uma pista oposta nesse sentido, com baixa carga aerodinâmica e longas retas, ela exhibe a assinatura mais forte de composto e vento, cada detalhe na velocidade do vento e também na pressão atmosférica fazem

diferença e, além disso, a escolha do composto correto também. Bahrein e Estados Unidos são governados pela temperatura de pista. Nessa parte, estamos tratando de uma corrida que ocorre no deserto, mesmo que a noite, e outra que ocorre na parte da tarde, no calor abrasivo do Texas. A Arábia Saudita é o único circuito em que o efeito de equipe rivaliza com o de clima e pneu, o que evidencia a grande dominância que a Red Bull Racing teve no circuito durante o regulamento.

Do ponto de vista estratégico, essas diferenças sugerem prioridades distintas de decisão em cada Grande Prêmio, respondendo à **QP4**. Em Hungria e Itália, onde a vida do pneu e o composto lideram a componente não linear do resíduo, o modelo recomenda concentrar o esforço de simulação na janela de parada e na escolha do composto, pois é ali que o tempo de volta responde de maneira mais acentuada e menos previsível linearmente. Em Bahrein e Estados Unidos, a alavanca principal é térmica: a preparação deve priorizar o gerenciamento de temperatura de pista e o ajuste do carro à faixa térmica ideal. Na Arábia Saudita, o deslocamento da importância da Red Bull Racing entre a validação e o *holdout* mostra que o modelo é capaz de detectar mudanças de hierarquia competitiva entre temporadas, um sinal útil para calibrar expectativas de referência de ritmo. Por fim, o perfil de recência oferece uma orientação improtante: janelas longas de contexto, como as 17 voltas selecionadas para o Bahrein, adicionam parâmetros sem adicionar informação proporcional à essa carga. É necessário avaliar cada cenário, mas para a configuração do LSTM, é importante validar o tamanho dessa janela para cada contexto de uma maneira mais eficiente.

4.5 Considerações Finais

Os resultados apresentados neste capítulo respondem, em conjunto, às cinco questões de pesquisa que orientaram este trabalho. As Seções 4.1 e 4.2 mostraram que a agregação multi-piloto e o modelo híbrido MLR-LSTM melhoram a predição do tempo de volta em cenários de dados limitados, com ganho estatisticamente significativo na Itália e desempenho competitivo nos demais circuitos, ao passo que o indicador COS permitiu decidir entre modelos empatados estatisticamente. As Seções 4.3 e 4.4, por sua vez, mostraram que essa predição não é uma caixa-preta: os coeficientes da MLR, o ganho e os valores SHAP do XGBoost, e a extensão do SHAP ao resíduo do LSTM revelaram que cada circuito possui uma assinatura própria de fatores climáticos, mecânicos e temporais, e que o componente não linear do híbrido recupera especificamente aquilo que a formulação linear não consegue capturar, como o efeito de composto e degradação em Monza. Essa combinação entre desempenho preditivo e interpretabilidade é o que permite traduzir os resultados em *insights* estratégicos específicos para cada Grande Prêmio, e não apenas em métricas de erro.

Tabela 9 – Síntese dos *insights* estratégicos por circuito.

Circuito	Fatores Climáticos	Fatores Físicos	Fatores Temporais
Bahrein	Temperatura de pista é o fator climático dominante nos três modelos; 2º maior ganho no XGBoost). No <i>holdout</i> , pressão e vento assumem a frente e o resíduo do LSTM passa a ser liderado pela umidade.	<i>TyreLife</i> aparece nos três modelos e o composto duro C1 figura no ganho. Equipes de menor desempenho (Kick Sauber, Haas) aparecem aumentando o tempo de volta.	Forte autorregressão e queda do tempo com o número da volta (-0.176). A busca selecionou 17 voltas de contexto, mas a volta alvo concentra 34.3% do <i>SHAP</i> médio e as onze voltas mais antigas somam menos de 18%: memória efetiva curta, janela superdimensionada.
Arábia Saudita	Umidade é o clima determinante e o único que <i>cresce</i> no <i>holdout</i> . Velocidade do vento e delta de temperatura dominam o resíduo do LSTM em 2025 (0.0139 e 0.0094).	Maior efeito de equipe do conjunto: a Red Bull Racing é a 2ª maior importância no ganho, no <i>holdout</i> , no entanto, cai em impacto no SHAP, sinalizando o fim de sua dominância no traçado em 2025. Composto C3 relevante no resíduo.	Tempo anterior e número da volta concentram a explicação. Sequência ótima de apenas 2 voltas, com 56.7% da magnitude na volta alvo.
Estados Unidos	Temperatura de pista é o 2º fator em todos os modelos e o de maior coeficiente negativo. A pressão é o 2º coeficiente positivo da MLR, mas tem ganho baixo (11ª de 50) no XGBoost. Umidade cresce em 2025.	A vida do pneu não aparece importar tanto frente às demais pistas; existe uma expressão mais por piloto e equipe (Albon, Red Bull, Haas no ganho) do que por degradação.	Tempo de volta anterior dominante, mas o número da volta é o <i>maior</i> canal do resíduo do LSTM (0.021): a evolução da pista ao longo da corrida escapa à MLR. Recência mais distribuída do conjunto (27.4% na volta alvo, com 6 voltas de contexto).
Itália	Clima secundário na validação (delta de temperatura, umidade e pressão no resíduo), mas o vento assume a 3ª posição no <i>holdout</i> (0.0149), coerente com um traçado de baixa carga aerodinâmica e longas retas.	Composto C4 e <i>TyreLife</i> lideram o resíduo do LSTM, embora a MLR praticamente os ignore (45ª e 46ª posições em coeficiente). É o efeito não linear que o híbrido recupera.	Único circuito em que o Ano supera o tempo anterior no ganho, revelando forte evolução entre temporadas. Memória curta (3 voltas; 48.3% na volta alvo).
Hungria	Pressão atmosférica é o traço climático mais forte de todo o conjunto, com o 2º maior ganho e 2º maior SHAP no <i>holdout</i> (0.145), seguida da temperatura de pista.	Degradação elevada: <i>TyreLife</i> é o 2º coeficiente positivo da MLR e o 2º canal do resíduo do LSTM, com os compostos C3 e C4 logo atrás. Faz sentindo real, visto que é um traçado de baixa velocidade e alto desgaste.	Tempo anterior com ganho alto e número da volta como maior canal do resíduo. Memória mais curta do conjunto: 59.1% da magnitude na volta alvo, com 3 voltas de contexto.

O Capítulo 5 discute os objetivos deste trabalho, além de abordar as limitações encontradas durante os resultados de forma sistemática, como os ganhos limitados do modelo híbrido e otimização das sequências. Além disso, os resultados ressaltam que os *insights* extraídos dizem respeito ao ritmo puro de corrida, e não levam em conta fatores aleatórios. Esse capítulo também discute as contribuições que os resultados apresentados permitem reivindicar para a literatura e aponta os caminhos que a extensão deste trabalho deve seguir.

5 Conclusões e Trabalhos Futuros

Este capítulo encerra a monografia, retomando o que foi desenvolvido ao longo dos capítulos anteriores e apontando os caminhos que a continuidade deste trabalho deve seguir. A Seção 5.1 sintetiza as cinco etapas da metodologia empregada, detalhada no Capítulo 3, consolidando em um só lugar o *pipeline* que sustenta os resultados obtidos. A Seção 5.2 avalia se os objetivos definidos foram alcançados e discute os resultados apresentando no Capítulo 4. A Seção 5.3 discute, de forma crítica, as limitações encontradas pelo trabalho e os aspectos que os resultados ainda não são capazes de explicar. A Seção 5.4 reúne as contribuições deste trabalho para a literatura, organizadas segundo as cinco questões de pesquisa que o orientaram. Por fim, a Seção 5.5 conecta os resultados obtidos a três frentes de continuidade para trabalhos futuros.

5.1 Síntese da Metodologia Utilizada

A metodologia empregada neste trabalho organizou-se em cinco etapas principais, detalhadas no Capítulo 3. A primeira etapa consistiu na extração e no pré-processamento de dados públicos de telemetria da biblioteca FastF1, restritos à corrida principal de cinco circuitos entre as temporadas de 2022 e 2025, com filtragem de *outliers* por percentil e mapeamento dos compostos de pneu Pirelli por meio de *web scraping*, de modo a assegurar que os dados modelados representassem o ritmo puro de corrida. A segunda etapa definiu a configuração experimental, empregando validação temporal por janela expansiva nos *baselines* para evitar vazamento de dados futuros. A terceira etapa consolidou os modelos de Regressão Linear Múltipla e XGBoost como *baselines* e um modelo híbrido MLR-LSTM com aprendizado residual em dois estágios. A quarta etapa introduziu o indicador COS para selecionar o modelo mais adequado em cenários de empate estatístico entre as métricas tradicionais. A quinta e última etapa apresentou o protocolo de interpretabilidade referente às três famílias de modelos, introduzindo coeficientes de regressão, importância de *features* e valores SHAP para o XGBoost, e valores SHAP obtidos via *GradientExplainer* para o componente residual do LSTM.

5.2 Discussão sobre as Questões de Pesquisa

O objetivo geral deste trabalho, desenvolver um *pipeline* de aprendizado de máquina capaz de prever tempos de volta com precisão em cenários de dados limitados e de extrair *insights* estratégicos interpretáveis a partir dessa predição, foi alcançado, ainda que com detalhes importantes que são detalhados na Seção 5.3. Os cinco objetivos específicos,

derivados das questões de pesquisa apresentadas, foram atendidos com as seguintes ponderações.

A **QP1** propunha investigar estratégias de predição do tempo de volta em cenários de dados limitados. Este objetivo foi alcançado por meio da estratégia de agregação multi-piloto por circuito. A Tabela 7 demonstra que a filtragem individual por piloto produz modelos com $R^2 < 0$ em todos os circuitos, ao passo que a série agregada (Tabela 8) eleva o desempenho a valores de R^2 entre 0.84 e 0.92 no conjunto de teste, confirmando a viabilidade da abordagem proposta.

A **QP2** propunha desenvolver uma estratégia para melhorar a acurácia da predição do tempo de volta de pilotos. Este objetivo foi parcialmente alcançado: o modelo Híbrido MLR-LSTM obteve o melhor desempenho de treino em todos os cinco circuitos e superioridade estatisticamente significativa no conjunto de teste apenas no Grande Prêmio da Itália, com empate estatístico nos demais circuitos. O objetivo foi, portanto, cumprido no sentido de que a arquitetura híbrida se mostrou uma estratégia válida e, em ao menos um cenário, superior aos *baselines*, mas não no sentido de uma superioridade generalizada a todos os circuitos analisados. É importante ressaltar que análise extensa realizar na seção de resultados demonstra que, o modelo híbrido consegue ser superior em outros cenários, como o contexto específico de uma corrida nos Estados Unidos.

O terceiro objetivo, referente à **QP3**, propunha um critério para selecionar o modelo mais adequado equilibrando desempenho médio e variância. Este objetivo foi alcançado com a proposição do indicador COS, cuja utilidade foi demonstrada justamente nos circuitos em que as métricas tradicionais não permitiam decisão, como XGBoost para Arábia Saudita. Além disso, o indicador aponta diretamente para quedas grandes de desempenho em treino *versus* teste.

O quarto e último objetivo, referente à **QP4**, se concentrava em identificar como cada *feature*, seja ela climática, mecânica ou temporal mais influenciavam a saída dos modelos em cada circuito. Este objetivo foi alcançado por meio do protocolo de interpretabilidade transversal às três famílias de modelos, detalhado na Seção 3.5 da Metodologia e aplicado na Seção 4.3 dos Resultados. Além disso, de maneira mais completa, esse objetivo propunha identificar como as informações de explicabilidade encontradas poderiam construir *insights* estratégicos referentes à cada circuito. O alcance desse objetivo pode ser visto por meio da Tabela 9, que consolida, por circuito, as leituras complementares da MLR, do XGBoost e do componente residual do modelo híbrido, construindo o perfil de cada pista. Aqui, ocorre uma das partes altas desse trabalho: destrinchando a contribuição do componente LSTM na composição do modelo híbrido é possível entender quais fatores a abordagem linear deixa escapar. Na Itália, caso de maior sucesso, essa resposta é bem clara, identificando a força da vida do pneu e a escolha do composto certo.

5.3 Limitações do Trabalho

Os resultados obtidos por esse trabalho estão sujeitos a determinadas limitações em sua conclusão. Primeiramente, os ganhos do modelo híbrido são pequenos e inconsistentes: ele supera o próprio baseline de maneira significativa apenas na Itália, enquanto na Arábia Saudita consegue superar em termos de médias. Ainda assim, no contexto da Fórmula 1, pequenas variações de décimos ou até milésimos de segundo podem ter relevância prática, o que torna importante avaliar não apenas a significância estatística, mas também o impacto operacional desses ganhos. A busca de hiperparâmetros no cenário do aprendizado profundo também apresentou sinais de instabilidade, como o comprimento de sequência de 17 voltas selecionado para o Bahrein. Através da análise do perfil de recência, é possível observar que informações de voltas posteriores não agregam tanto impacto na saída do modelo. Na verdade, a maioria das sequências encontradas para o LSTM em cada uma das pistas é normalmente curta, como, por exemplo, para Itália, onde temos o melhor desempenho e uma sequência ótima de 3 voltas. O comprimento da sequência foi encontrado através do menor RMSE de validação, portanto outro tipo de análise talvez seja necessária, como uma análise exploratória mais aprofundada da natureza dos *stints* em cada uma das pistas.

Há também limitações de dados e de interpretação. Enquanto esse trabalho foca exclusivamente em extrair informações do ritmo de corrida puro, os *insights* estratégicos derivados se aplicam apenas ao ritmo em condições de pista limpa. Portanto, limitam-se em explicar o comportamento dos carros e da corrida nesse cenário de alto desempenho, oferecendo *insights* que auxiliam na formação da estratégia ótima, mas que não podem caracterizá-la como um todo, visto que excluí os cenários de aleatoriedade presentes, como: *Safety Car* e chuva. De outra maneira, o bloco de teste corresponde a uma única corrida de 2025 por circuito, entre 460 e 880 voltas, o que leva a uma avaliação em condições específicas daquele fim de semana. A validação, principalmente na MLR e no XGBoost, ocorre ao longo dos anos, através das janelas expansivas. Portanto, as mesmas podem trazer explicações mais consistentes sobre o caráter de cada pista. Por fim, os modelos foram treinados de forma independente por circuito, sobre cinco traçados, sem qualquer mecanismo de transferência entre eles, o que impede generalizar as conclusões para pistas não estudadas.

5.4 Principais Contribuições

Ao endereçar as cinco questões de pesquisa, este trabalho apresenta as seguintes contribuições:

1. **Agregação multi-piloto por circuito como estratégia para escassez de dados (QP1).** Propõe-se organizar os tempos de volta de *todos* os pilotos de um

dado circuito em uma única série temporal combinada, expandida ao longo de múltiplas temporadas, em vez de uma série por piloto. A Tabela 7 evidencia o custo da abordagem individual, enquanto a série agregada eleva o desempenho a R^2 entre 0.84 e 0.92 no conjunto de teste, com reduções de MAE de mais de uma ordem de grandeza em circuitos como os Estados Unidos (1.92 seg. para 0.26 seg.);

2. **Modelo híbrido MLR–LSTM por correção de resíduo (QP2).** Propõe-se uma arquitetura em dois estágios na qual uma rede LSTM aprende o resíduo de um *baseline* tabular, sendo a predição final a soma de ambos. O híbrido obtém o melhor desempenho de treino em todos os cinco circuitos e é o único modelo estatisticamente superior no Grande Prêmio da Itália, onde é ao menos 11% e 27% mais preciso que a MLR e o XGBoost, respectivamente, com intervalos de confiança que não se sobrepõem;
3. **COS para seleção de modelos sob empate estatístico (QP3).** Introduz-se uma métrica indicativa que combina desempenho médio e variabilidade do modelo, razão entre o comportamento no teste e no treino. Sua utilidade prática é demonstrada justamente nos circuitos em que os intervalos de confiança se sobrepõem e as métricas tradicionais não decidem: nesses casos, o COS opera como critério de desempate, apontando a MLR para Bahrein e Hungria e o XGBoost para a Arábia Saudita;
4. **Protocolo de interpretabilidade transversal às três famílias de modelos (QP4).** Este trabalho estende a interpretabilidade para além do *baseline*, articulando coeficientes padronizados da MLR, ganho e valores SHAP do XGBoost e valores SHAP do resíduo do LSTM, estes últimos obtidos por GradientExplainer sobre as sequências tridimensionais, sem retreinamento. Deriva-se ainda um *perfil de recência*, que quantifica a magnitude média dos valores SHAP por posição da sequência e permite diagnosticar quanta memória o modelo de fato utiliza, revelando, por exemplo, que a janela de 17 voltas selecionada para o Bahrein é subaproveitada, com menos de 18% da atribuição nas onze voltas mais antigas;
5. **Caracterização da assinatura de cada circuito e extração de *insights* estratégicos (QP4).** O cruzamento das leituras da explicabilidade dos três modelos produz a Tabela 9, que separa, por pista, os fatores climáticos, físicos e temporais determinantes. A partir dessas informações, é possível extrair uma análise estratégica única para cada circuito.

5.5 Trabalhos Futuros

Os resultados obtidos abrem três frentes principais de continuidade. A primeira diz respeito ao aprofundamento da explicabilidade. A análise conduzida neste trabalho

revelou uma limitação de atribuição no modelo híbrido: como Piloto e Equipe entram no LSTM por camadas de *embedding* com códigos inteiros, cujos gradientes são nulos, o *GradientExplainer* não é capaz de atribuir importância a essas variáveis, e a interpretação do resíduo restringe-se aos canais densos. Pretende-se contornar essa restrição com o uso de Local Interpretable Model-Agnostic Explanations (LIME). Em paralelo, pretende-se incorporar uma camada de atenção à arquitetura (KISER; SHI; BUCHER, 2024). Além do ganho potencial de acurácia, a atenção tornaria o peso atribuído a cada volta passada um parâmetro explícito do modelo, legível diretamente, e não uma quantidade inferida posteriormente, como no perfil de recência aqui proposto.

A segunda frente trata da otimização da busca de hiperparâmetros, com ênfase no LSTM e, sobretudo, na definição do contexto sequencial. O perfil de recência evidenciou que a busca atual pode convergir para configurações que o próprio modelo não aproveita: no Bahrein, foram selecionadas 17 voltas de contexto, mas menos de 18% da magnitude média dos valores SHAP recai sobre as onze voltas mais antigas, enquanto a volta alvo concentra 34.3%. Trata-se de um custo computacional e de um acréscimo de parâmetros sem retorno informacional proporcional. Pretende-se, portanto, reformular a seleção do comprimento de sequência, seja restringindo o espaço de busca com base em evidência de recência obtida previamente, seja incorporando ao critério de otimização uma penalização explícita ao tamanho da janela, seja ainda avaliando estratégias de busca mais eficientes que a atualmente empregada. O objetivo é que o comprimento da sequência deixe de ser um hiperparâmetro escolhido às cegas e passe a refletir a memória efetiva do fenômeno em cada circuito, que este trabalho mostrou variar de duas a dezessete voltas entre as pistas analisadas.

A terceira frente, de caráter aplicado, é o desenvolvimento de uma simulação de corrida que integre o melhor modelo identificado para cada circuito, segundo o critério do COS. A proposta é transformar o preditor em uma ferramenta de apoio à decisão, capaz de projetar o ritmo restante sob diferentes cenários de estratégia, como variações na janela de parada e na escolha do composto. É nesse ponto que a análise de interpretabilidade deixa de ser um resultado terminal e passa a ser insumo de projeto: a Tabela 9 indica, para cada traçado, quais fatores o modelo considera determinantes, e essa informação orienta diretamente quais variáveis a simulação deve permitir manipular. Em Hungria e Itália, onde a vida do pneu e o composto lideram a componente não linear do resíduo, o esforço de simulação se concentraria na estratégia de pneus; em Bahrein e Estados Unidos, governados pela temperatura de pista, o cenário térmico seria a variável central. Cabe registrar que a viabilidade dessa simulação depende de um passo adicional de tratamento de dados, uma vez que o pré-processamento atual descarta voltas de entrada e saída de *box* e períodos de *safety car*, precisamente os eventos que uma simulação estratégica precisa representar.

Referências

- ADWAN, I. et al. Asphalt pavement temperature prediction models: A review. *Applied Sciences*, MDPI, v. 11, n. 9, p. 3794, 2021. Disponível em: <<https://doi.org/10.3390/app11093794>>. Citado na página 39.
- BANSAL, A. et al. *Advanced Machine Learning Approaches for Formula 1 Race Performance Prediction: A Comprehensive Analysis of Championship Point Forecasting*. 2025. Preprint available at ResearchGate. Citado 3 vezes nas páginas 16, 31 e 33.
- BEKKER, J.; LOTZ, W. Planning formula one race strategies using discrete-event simulation. *Journal of the Operational Research Society*, Springer, v. 60, n. 7, p. 952–961, 2009. Citado 2 vezes nas páginas 30 e 33.
- CHETELES, O.-A. *Feature Importance versus Feature Selection in Predictive Modeling for Formula 1 Race Standings*. Dissertação (Mestrado) — University of Twente, Enschede, July 2024. Citado 3 vezes nas páginas 17, 30 e 33.
- FATIMA, S. S. W.; JOHRENDT, J. Deep-racing: An embedded deep neural network (ednn) model to predict the winning strategy in formula one racing. *International Journal of Machine Learning*, v. 13, n. 3, 2023. Citado 5 vezes nas páginas 16, 17, 31, 34 e 48.
- Fédération Internationale de l'Automobile. *F1's New Era: Everything you need to know about how the FIA is making Formula 1 more competitive, more sustainable and safer in 2026*. 2025. <<https://www.fia.com/news/fls-new-era-everything-you-need-know-about-how-fia-making-formula-1-more-competitive-more>>. Accessed: 2026-05-03. Citado 2 vezes nas páginas 16 e 23.
- FONG, S. J. et al. Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. *International Journal of Interactive Multimedia and Artificial Intelligence*, v. 6, n. 1, p. 132–140, Mar. 2020. Citado na página 47.
- GARCÍA, M. P. Q. *Optimization of Pit Stop Strategies in Formula 1 Racing: A Data-Driven Approach*. Master's Thesis — University of Chile, 2024. Citado 4 vezes nas páginas 16, 17, 30 e 33.
- GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. ed. Sebastopol, CA: O'Reilly Media, 2019. Citado na página 43.
- HABER, E. E. et al. Formula 1 Race Winner Prediction Using Random Forest and SHAP Analysis. In: *2025 International Conference on Control, Automation, and Instrumentation (IC2AI)*. Beirut, France: IEEE, 2025. p. 1270–1274. Citado 4 vezes nas páginas 31, 32, 34 e 35.
- HEILMEIER, A. et al. Application of monte carlo methods to consider probabilistic effects in a race simulation for circuit motorsport. *Applied Sciences*, Multidisciplinary Digital Publishing Institute, v. 10, n. 12, p. 4229, 2020. Citado 3 vezes nas páginas 17, 30 e 33.

- HEILMEIER, A.; GRAF, M.; LIENKAMP, M. A race simulation for strategy decisions in circuit motorsports. In: IEEE. *Proc. of the IEEE 21st ITSC*. [S.l.], 2018. p. 2986–2993. Citado 3 vezes nas páginas 17, 30 e 33.
- HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 3rd. ed. Melb., Australia: OTexts, 2021. Citado na página 45.
- JAFRI, A. *Predicting Formula 1 Race Outcomes: A Machine Learning Approach*. 2024. Capstone Project, New York University Abu Dhabi. Citado na página 34.
- KARA, M. F.; KÜÇÜKMANISA, A. Predicting lap times in formula 1: A deep learning approach using lstm networks. In: *2025 IEEE 21st ICCP*. [S.l.: s.n.], 2025. p. 1–5. Citado 3 vezes nas páginas 18, 31 e 34.
- KISER, A. C.; SHI, J.; BUCHER, B. T. An explainable long short-term memory network for surgical site infection identification. *Surgery*, v. 176, p. 24–31, 2024. Citado na página 72.
- KRZYSZTOŃ, S. A.; SMOŁKA, J. Application of machine learning for predicting formula 1 race results. *Journal of Computer Sciences Institute*, 2026. Citado 3 vezes nas páginas 16, 31 e 33.
- LIU, X.; FOTOUHI, A. Formula-e race strategy development using artificial neural networks and monte carlo tree search. *Neural Computing and Applications*, Springer, v. 32, n. 18, p. 15191–15207, 2020. Citado 3 vezes nas páginas 17, 32 e 33.
- MAHMUD, A. et al. Hybrid arima-lstm for covid-19 forecasting: a comparative ai modeling study. *PeerJ Comp. Sci.*, v. 11, p. e3195, Sep. 2025. Citado na página 47.
- MICHELUCCI, U.; VENTURINI, F. *Best Practices for Machine Learning Experimentation in Scientific Applications*. 2025. Disponível em: <<https://arxiv.org/abs/2511.21354>>. Citado na página 49.
- PEIXEIRO, M. *Time Series Forecasting in Python*. [S.l.]: Manning Publications, 2022. ISBN 9781617299889. Citado na página 45.
- PICCINOTTI, D. *Open Loop Planning for Formula 1 Race Strategy identification*. Master's Thesis — Politecnico di Milano, 2021. Citado 2 vezes nas páginas 32 e 33.
- SASIKUMAR, A.; LEEMA, A. A.; BALAKRISHNAN, P. Data-driven pit stop decision support for formula 1 using deep learning models. *Frontiers in Artificial Intelligence*, Volume 8, 2025. Citado 2 vezes nas páginas 31 e 34.
- STEPIEN, Z. The pro-ecological evolution of powertrains and fuels in formula 1. *Energies*, MDPI, v. 18, n. 22, p. 6013, 2025. Disponível em: <<https://doi.org/10.3390/en18226013>>. Citado na página 16.
- THOMAS, D. et al. Explainable reinforcement learning for formula one race strategy. *arXiv preprint arXiv:2501.04068*, 2025. Disponível em: <<https://arxiv.org/abs/2501.04068>>. Citado 4 vezes nas páginas 16, 17, 32 e 33.
- TODD, J. et al. Explainable time series prediction of tyre energy in formula one race strategy. In: *Proc. 40th SIGAPP*. [S.l.]: ACM, 2025. (SAC '25), p. 1082–1089. Citado 3 vezes nas páginas 17, 31 e 34.

VATSA, A. et al. Residual lstm-based short duration forecasting of polarization current for effective assessment of transformers insulation. *Sci. Rep.*, v. 14, n. 1369, 2024. Citado na página 47.

ZHAO, Z. Deep neural network-based lap time forecasting of formula 1 racing. *Applied and Computational Engineering*, v. 47, p. 61–66, 03 2024. Citado 3 vezes nas páginas 18, 31 e 34.

Apêndices

APÊNDICE A – Análise Exploratória da Característica das Pistas

Nessa seção, serão mostradas as tabelas referentes a caracterização de cada um dos circuitos estudados, contendo: sumarização estatística e média do tempo de volta ao longo dos anos.

A.1 Sumarização Estatística para os Dados Filtrados

Tabela 10 – Tabela de sumarização estatística dos dados para a pista de Bahrein.

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_prev	98.16	1.58	0.02	94.09	97.17	98.09	99.10	103.77	0.24	0.22
TyreLife	11.38	5.65	0.50	3.00	7.00	11.00	15.00	31.00	0.68	0.06
LapNumber	29.24	16.19	0.55	3.00	16.00	28.00	43.00	57.00	0.06	-1.21
LapTime_sec	98.23	1.58	0.02	94.15	97.26	98.16	99.18	104.45	0.23	0.20
Humidity	38.54	11.32	0.29	20.00	29.00	44.00	49.00	51.00	-0.56	-1.24
Pressure	1013.47	4.00	0.00	1008.00	1010.20	1016.90	1017.20	1017.50	-0.22	-1.81
TrackTemp	27.60	3.35	0.12	21.90	23.50	28.70	30.50	32.00	-0.51	-1.24
WindSpeed	0.76	0.69	0.91	0.00	0.30	0.60	1.00	2.70	1.10	0.09
TempDelta	4.15	0.75	0.18	2.40	3.67	4.30	4.70	5.50	-0.51	-0.60

Tabela 11 – Tabela de sumarização estatística dos dados para a pista de Arábia Saudita.

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_prev	94.64	1.39	0.01	91.63	93.57	94.62	95.64	99.15	0.27	-0.40
TyreLife	17.66	10.07	0.57	3.00	9.00	16.00	25.00	49.00	0.58	-0.54
LapNumber	27.20	13.74	0.51	3.00	14.00	28.00	39.00	50.00	-0.10	-1.19
LapTime_sec	94.56	1.36	0.01	91.63	93.52	94.52	95.56	98.66	0.23	-0.53
Humidity	58.42	5.09	0.09	47.00	55.00	59.00	62.00	67.00	-0.52	-0.32
Pressure	1010.90	2.46	0.00	1006.70	1007.30	1011.20	1013.00	1013.40	-0.73	-1.12
TrackTemp	31.95	3.40	0.11	27.30	30.20	30.90	36.20	38.30	0.63	-0.93
WindSpeed	1.28	0.63	0.49	0.00	0.80	1.20	1.60	4.00	1.10	1.75
TempDelta	5.00	1.79	0.36	1.90	4.20	5.00	5.60	9.40	0.33	-0.13

Tabela 12 – Tabela de sumarização estatística dos dados para a pista de Hungria.

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_prev	83.86	1.54	0.02	79.41	82.88	84.02	85.00	89.33	-0.36	-0.26
TyreLife	15.65	8.59	0.55	3.00	9.00	14.00	21.00	55.00	0.72	0.19
LapNumber	35.33	19.37	0.55	3.00	18.00	35.00	52.00	70.00	0.06	-1.21
LapTime_sec	83.88	1.53	0.02	79.41	82.91	84.04	85.03	89.84	-0.37	-0.25
Humidity	51.36	13.74	0.27	31.00	43.00	55.00	58.00	72.00	-0.00	-1.23
Pressure	985.60	2.53	0.00	981.40	981.70	986.90	987.30	988.50	-0.73	-1.03
TrackTemp	37.74	8.97	0.24	23.70	30.70	32.10	46.70	50.10	-0.02	-1.72
WindSpeed	2.17	1.28	0.59	0.30	1.20	1.70	3.00	5.80	0.90	-0.23
TempDelta	13.00	4.51	0.35	5.70	9.10	10.10	17.40	21.30	0.28	-1.52

Tabela 13 – Tabela de sumarização estatística dos dados para a pista de Itália.

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_prev	85.42	1.76	0.02	81.00	83.92	85.60	86.97	89.85	-0.26	-0.98
TyreLife	15.38	9.11	0.59	3.00	8.00	14.00	21.00	50.00	0.92	0.48
LapNumber	26.38	14.18	0.54	3.00	14.00	26.00	38.00	53.00	0.07	-1.16
LapTime_sec	85.39	1.76	0.02	80.90	83.88	85.59	86.96	89.64	-0.26	-1.01
Humidity	36.62	5.03	0.14	30.00	31.75	39.00	41.00	44.00	-0.06	-1.84
Pressure	995.83	2.99	0.00	992.50	993.00	996.30	1000.20	1000.50	0.44	-1.38
TrackTemp	43.82	1.98	0.05	40.00	42.80	43.60	44.40	50.70	1.13	2.04
WindSpeed	1.81	0.66	0.36	0.60	1.30	1.70	2.20	4.00	0.92	0.87
TempDelta	14.36	1.62	0.11	10.80	13.20	14.20	15.40	17.90	0.19	-0.65

Tabela 14 – Tabela de sumarização estatística dos dados para a pista de Estados Unidos.

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_prev	101.25	1.74	0.02	97.61	99.90	101.16	102.54	106.83	0.23	-0.66
TyreLife	14.41	7.60	0.53	3.00	8.00	14.00	19.00	40.00	0.55	-0.29
LapNumber	30.30	15.55	0.51	3.00	16.00	31.00	44.00	56.00	-0.07	-1.24
LapTime_sec	101.24	1.73	0.02	97.61	99.90	101.16	102.53	106.12	0.21	-0.69
Humidity	34.51	9.88	0.29	20.00	21.00	36.00	41.00	51.00	-0.13	-1.15
Pressure	998.06	6.00	0.01	989.70	992.90	1000.50	1005.50	1006.60	0.07	-1.56
TrackTemp	42.97	4.10	0.10	34.60	40.70	45.40	46.27	48.00	-0.59	-1.04
WindSpeed	2.88	1.87	0.65	0.00	1.40	2.35	4.40	8.60	0.82	-0.28
TempDelta	13.57	4.66	0.34	5.60	10.30	16.40	17.40	19.90	-0.35	-1.44

A.2 Sumarização Estatística para os Dados Brutos

Tabela 15 – Tabela de sumarização estatística dos dados brutos para a pista de Bahrein (voltas: 4438 registros; clima: 639 registros).

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_sec	100.56	7.48	0.07	92.61	97.48	98.50	100.01	149.74	3.48	13.05
TyreLife	10.27	6.14	0.60	1.00	5.00	10.00	14.00	31.00	0.57	-0.22
LapNumber	28.53	16.32	0.57	1.00	14.00	28.00	42.75	57.00	0.02	-1.19
AirTemp	24.10	3.76	0.16	17.60	22.30	25.70	27.00	29.80	-0.60	-1.12
Humidity	36.14	12.04	0.33	16.00	22.00	38.00	46.00	51.00	-0.33	-1.56
Pressure	1013.13	3.97	0.00	1007.80	1010.00	1010.70	1017.00	1017.50	-0.11	-1.83
TrackTemp	28.80	3.55	0.12	21.90	26.30	29.80	31.30	35.10	-0.44	-0.88
WindSpeed	0.96	0.81	0.84	0.00	0.40	0.60	1.50	3.60	0.98	0.03

Tabela 16 – Tabela de sumarização estatística dos dados brutos para a pista de Arábia Saudita (voltas: 3562 registros; clima: 588 registros).

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_sec	97.09	8.92	0.09	91.63	93.76	94.97	96.19	165.66	4.08	17.07
TyreLife	15.63	10.79	0.69	1.00	7.00	14.00	23.00	49.00	0.63	-0.48
LapNumber	24.89	14.37	0.58	1.00	12.00	25.00	37.00	50.00	0.04	-1.19
AirTemp	26.60	1.74	0.07	25.20	25.50	25.80	26.40	31.70	1.51	1.07
Humidity	60.60	5.48	0.09	47.00	57.00	60.00	64.00	74.00	0.08	0.33
Pressure	1010.84	2.44	0.00	1006.30	1010.50	1012.00	1012.70	1013.50	-0.88	-0.81
TrackTemp	32.32	3.53	0.11	27.20	29.70	31.50	34.10	39.40	0.57	-0.86
WindSpeed	1.63	0.79	0.48	0.00	1.00	1.50	2.20	4.00	0.55	-0.51

Tabela 17 – Tabela de sumarização estatística dos dados brutos para a pista de Hungria (voltas: 5358 registros; clima: 643 registros).

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_sec	85.05	4.55	0.05	79.41	83.10	84.26	85.47	124.98	3.75	16.60
TyreLife	14.92	9.40	0.63	1.00	7.00	14.00	21.00	55.00	0.65	0.02
LapNumber	34.97	20.03	0.57	1.00	18.00	35.00	52.00	70.00	0.02	-1.20
AirTemp	24.51	4.49	0.18	17.80	19.30	22.40	28.90	30.30	-0.19	-1.67
Humidity	51.66	13.69	0.26	31.00	34.00	55.00	66.00	72.00	-0.10	-1.29
Pressure	985.74	2.52	0.00	981.40	985.20	986.20	987.70	989.00	-0.66	-0.96
TrackTemp	38.15	9.13	0.24	22.80	28.60	37.90	47.10	51.40	-0.07	-1.65
WindSpeed	2.24	1.25	0.56	0.30	1.20	1.80	3.20	5.80	0.72	-0.56

Tabela 18 – Tabela de sumarização estatística dos dados brutos para a pista de Itália (voltas: 3912 registros; clima: 580 registros).

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_sec	87.22	6.89	0.08	80.90	84.26	86.09	87.34	145.33	4.44	23.74
TyreLife	14.38	9.66	0.67	1.00	7.00	13.00	20.00	51.00	0.82	0.25
LapNumber	26.20	15.09	0.58	1.00	13.00	26.00	39.00	53.00	0.04	-1.19
AirTemp	29.15	2.41	0.08	25.90	27.10	28.50	30.00	34.10	0.72	-0.80
Humidity	37.77	5.25	0.14	30.00	32.00	40.00	42.00	46.00	-0.11	-1.60
Pressure	996.10	3.07	0.00	992.50	993.20	996.20	1000.20	1001.20	0.44	-1.38
TrackTemp	44.53	3.31	0.07	39.30	42.80	43.45	44.52	54.60	1.78	2.30
WindSpeed	1.49	0.68	0.46	0.00	1.00	1.50	1.80	4.00	0.75	0.89

Tabela 19 – Tabela de sumarização estatística dos dados brutos para a pista de Estados Unidos (voltas: 4132 registros; clima: 647 registros).

Feature	$\bar{x}$	s	CV	Min	Q1	Q2	Q3	Max	γ_1	κ
LapTime_sec	103.22	7.21	0.07	97.33	100.12	101.55	103.16	149.84	3.86	16.14
TyreLife	13.04	8.31	0.64	1.00	6.00	12.00	19.00	40.00	0.57	-0.36
LapNumber	28.05	16.19	0.58	1.00	14.00	28.00	42.00	56.00	0.02	-1.21
AirTemp	28.94	1.06	0.04	26.40	28.40	28.90	29.70	31.00	-0.12	-0.50
Humidity	37.80	10.32	0.27	20.00	32.00	39.00	47.00	54.00	-0.34	-1.05
Pressure	997.88	6.09	0.01	989.70	992.70	994.90	1002.10	1008.00	0.28	-1.48
TrackTemp	41.77	4.45	0.11	33.40	37.70	41.60	45.80	48.00	-0.37	-1.22
WindSpeed	3.16	1.86	0.59	0.00	1.60	2.80	4.50	10.10	0.68	-0.29

A.3 Tempo de Volta ao Longo dos Anos

Tabela 20 – Média do tempo de volta (s) por piloto e por ano – Bahrein.

Piloto	2022	2023	2024	2025
VER	97.96	97.50	95.70	98.05
PER	98.21	97.68	96.05	–
LEC	97.66	98.33	96.27	97.71
HAM	98.28	98.40	96.47	98.01
SAI	98.06	98.32	96.10	98.97
RUS	98.73	98.42	96.42	98.20
SAR	–	99.00	97.30	–
PIA	–	101.15	96.52	97.20
NOR	100.79	98.33	96.48	97.59
GAS	99.59	98.33	97.34	98.04
ANT	–	–	–	98.37
ALO	99.63	98.19	96.80	98.94
TSU	99.50	98.60	97.51	98.49
MAG	99.06	99.12	97.39	–
BOT	99.39	98.53	97.69	–
HAD	–	–	–	98.63
DOO	–	–	–	98.63
BEA	–	–	–	98.66
OCO	99.38	99.30	97.73	98.33
ZHO	99.46	99.41	97.37	–
LAW	–	–	–	98.75
HUL	100.69	98.91	96.78	98.81
ALB	100.35	98.73	97.58	98.73
STR	100.64	98.42	97.14	99.23
RIC	100.37	–	97.43	–
BOR	–	–	–	99.31
DEV	–	99.34	–	–
MSC	100.10	–	–	–
LAT	100.86	–	–	–

Tabela 21 – Média do tempo de volta (s) por equipe e por ano – Bahrein.

Equipe	2022	2023	2024	2025
Red Bull Racing	98.09	97.59	95.88	98.30
Ferrari	97.85	98.32	96.19	97.85
Mercedes	98.52	98.41	96.45	98.30
McLaren	100.57	99.03	96.50	97.37
Alpine	99.50	98.67	97.52	98.32
Haas F1 Team	99.55	99.01	97.08	98.50
Racing Bulls	99.54	98.94	97.47	98.70
Kick Sauber	99.42	98.95	97.53	99.07
Aston Martin	100.66	98.31	96.98	99.09
Williams	100.61	98.86	97.47	98.83

Tabela 22 – Média do tempo de volta (s) por piloto e por ano – Arabia Saudita.

Piloto	2022	2023	2024	2025
PER	93.57	93.33	93.01	–
VER	93.58	93.72	92.84	93.15
LEC	93.59	94.50	93.27	93.36
ANT	–	–	–	93.85
RUS	94.47	94.16	93.70	93.76
PIA	–	95.52	93.52	93.19
SAI	93.68	94.31	–	94.49
BEA	–	–	93.77	94.69
HAM	95.16	94.44	93.65	93.84
NOR	95.35	95.57	93.47	93.26
ALO	95.63	93.85	93.60	94.62
LAW	–	–	–	94.45
HAD	–	–	–	94.50
MAG	95.09	95.10	94.76	–
GAS	95.36	94.86	–	–
OCO	95.28	94.85	95.00	95.43
HUL	96.20	95.22	94.11	95.17
TSU	–	95.35	95.07	–
ZHO	95.72	95.35	94.66	–
DOO	–	–	–	95.37
SAR	–	95.77	95.08	–
ALB	96.19	96.51	94.71	94.47
RIC	95.80	–	95.23	–
DEV	–	95.58	–	–
BOT	95.72	96.26	95.00	–
STR	96.21	95.61	95.70	95.28
BOR	–	–	–	95.83
LAT	98.09	–	–	–

Tabela 23 – Média do tempo de volta (s) por equipe e por ano – Arabia Saudita.

Equipe	2022	2023	2024	2025
Red Bull Racing	93.57	93.52	92.92	93.15
Ferrari	93.63	94.41	93.53	93.59
Mercedes	94.78	94.30	93.68	93.80
McLaren	95.52	95.54	93.50	93.22
Aston Martin	96.21	94.19	93.74	94.94
Haas F1 Team	95.09	95.16	94.43	95.06
Racing Bulls	95.36	95.47	95.15	94.48
Alpine	95.43	94.86	95.00	95.37
Kick Sauber	95.72	95.74	94.82	95.49
Williams	96.67	95.99	94.90	94.48

Tabela 24 – Média do tempo de volta (s) por piloto e por ano – Hungria.

Piloto	2022	2023	2024	2025
BOR	–	–	–	82.24
LAW	–	–	–	82.24
ANT	–	–	–	82.47
COL	–	–	–	82.51
HAD	–	–	–	82.58
PIA	–	84.32	83.30	81.00
VER	83.48	83.44	83.59	82.06
LEC	83.48	84.28	83.57	81.60
NOR	84.46	83.96	83.31	81.25
RUS	83.79	84.24	83.89	81.36
HAM	83.66	83.90	83.54	82.46
BEA	–	–	–	83.42
SAI	83.81	84.32	83.61	82.65
PER	83.57	83.86	83.70	–
ALO	85.06	84.36	84.46	82.06
STR	84.94	84.63	84.32	82.18
HUL	–	85.27	84.88	82.44
OCO	85.15	–	84.98	83.00
ALB	85.20	84.92	84.91	82.65
TSU	85.42	85.28	84.63	82.64
GAS	84.99	–	85.58	83.03
VET	84.81	–	–	–
RIC	85.03	84.89	84.73	–
SAR	–	85.23	84.95	–
BOT	85.32	84.88	85.09	–
MSC	85.12	–	–	–
ZHO	84.99	85.24	85.34	–
LAT	85.32	–	–	–
MAG	85.33	85.63	85.07	–

Tabela 25 – Média do tempo de volta (s) por equipe e por ano – Hungria.

Equipe	2022	2023	2024	2025
Red Bull Racing	83.53	83.65	83.64	82.33
McLaren	84.73	84.13	83.30	81.13
Mercedes	83.73	84.07	83.71	81.90
Ferrari	83.66	84.30	83.59	82.03
Aston Martin	84.88	84.49	84.39	82.12
Racing Bulls	85.16	85.09	84.67	82.40
Alpine	85.11	–	85.19	82.78
Kick Sauber	85.16	85.05	85.22	82.33
Williams	85.26	85.07	84.93	82.65
Haas F1 Team	85.22	85.44	84.98	83.18

Tabela 26 – Média do tempo de volta (s) por piloto e por ano – Italia.

Piloto	2022	2023	2024	2025
ANT	–	–	–	83.46
HAD	–	–	–	83.51
BEA	–	–	–	83.52
BOR	–	–	–	83.53
PIA	–	86.81	83.54	82.99
VER	85.36	86.13	84.18	82.50
COL	–	–	85.19	84.01
LEC	85.37	86.39	83.91	82.99
RUS	86.08	86.47	83.89	83.06
SAI	85.98	86.34	84.18	83.52
NOR	86.79	86.97	83.64	82.85
HAM	86.53	86.83	83.84	83.20
ALB	–	86.94	85.09	83.45
LAW	–	86.90	–	83.74
PER	86.17	86.27	84.49	–
TSU	87.36	–	86.20	83.88
GAS	87.23	87.01	85.26	84.06
ALO	87.67	86.92	84.64	84.50
HUL	–	87.23	84.82	–
STR	87.83	87.41	85.09	83.92
OCO	87.32	87.68	85.74	83.94
RIC	87.22	–	85.18	–
MAG	87.74	87.35	84.88	–
ZHO	87.28	87.07	85.97	–
BOT	87.36	87.26	85.93	–
DEV	87.30	–	–	–
SAR	–	87.33	–	–
MSC	87.42	–	–	–
LAT	87.52	–	–	–
VET	87.90	–	–	–

Tabela 27 – Média do tempo de volta (s) por equipe e por ano – Italia.

Equipe	2022	2023	2024	2025
Ferrari	85.66	86.36	84.04	83.09
Red Bull Racing	85.72	86.20	84.33	83.16
Mercedes	86.29	86.64	83.86	83.25
McLaren	86.99	86.89	83.58	82.92
Racing Bulls	87.30	86.90	85.20	83.62
Williams	87.40	87.13	85.14	83.49
Haas F1 Team	87.58	87.29	84.85	83.73
Aston Martin	87.84	87.14	84.82	84.08
Kick Sauber	87.32	87.17	85.95	83.53
Alpine	87.46	87.29	85.51	84.04

Tabela 28 – Média do tempo de volta (s) por piloto e por ano – Estados Unidos.

Piloto	2022	2023	2024	2025
ANT	–	–	–	100.02
VER	101.17	101.33	99.26	99.21
BEA	–	–	–	100.29
LAW	–	–	100.21	100.40
SAI	–	101.68	99.02	100.72
COL	–	–	100.11	100.86
HAD	–	–	–	100.51
BOR	–	–	–	100.56
LEC	101.88	102.15	98.90	99.32
PIA	–	103.04	99.36	99.67
NOR	102.91	101.53	99.20	99.26
HAM	101.56	101.35	–	99.62
RUS	102.07	101.71	99.85	99.81
HUL	–	102.71	99.95	100.05
PER	101.54	101.57	100.04	–
TSU	103.35	102.38	100.60	100.06
OCO	103.41	–	101.12	100.37
GAS	103.15	102.09	100.46	101.00
ALO	103.45	102.32	100.76	100.28
STR	103.78	101.95	100.88	100.32
ALB	103.73	102.68	100.92	100.25
MAG	103.90	102.97	100.08	–
ZHO	103.70	102.96	100.98	–
SAR	–	102.71	–	–
BOT	104.25	102.97	101.24	–
VET	102.84	–	–	–
RIC	103.82	103.32	–	–
MSC	103.70	–	–	–
LAT	104.37	–	–	–

Tabela 29 – Média do tempo de volta (s) por equipe e por ano – Estados Unidos.

Equipe	2022	2023	2024	2025
Red Bull Racing	101.35	101.45	99.66	99.63
Ferrari	101.88	101.91	98.96	99.47
Mercedes	101.81	101.52	99.85	99.92
McLaren	103.36	101.74	99.27	99.47
Aston Martin	103.03	102.12	100.82	100.30
Racing Bulls	103.25	102.81	100.39	100.45
Haas F1 Team	103.81	102.84	100.01	100.33
Alpine	103.43	102.09	100.76	100.93
Williams	104.05	102.69	100.48	100.27
Kick Sauber	103.81	102.96	101.10	100.30

APÊNDICE B – Resultados das Corridas em Cada Ano do Dataset

Essa seção apresenta os resultados finais para cada corrida presente no *dataset*, afim de trazer um contexto da realidade para os circuitos estudados. Em termos técnicos, **R** indicava que o piloto abandonou a corrida e **W** indica que o piloto se classificou mas não largou.

B.1 Resultados das Corridas

Tabela 30 – Posição de chegada de cada piloto por ano – Arabia Saudita.

Piloto	2022	2023	2024	2025
VER	1	2	1	2
PER	4	1	2	–
LEC	2	7	3	3
RUS	5	4	6	5
SAI	3	6	–	8
ANT	–	–	–	6
PIA	–	15	4	1
HAM	10	5	9	7
NOR	7	17	8	4
BEA	–	–	7	13
HAD	–	–	–	10
OCO	6	8	13	14
MAG	9	10	12	–
LAW	–	–	–	12
HUL	12	12	10	15
DEV	–	14	–	–
ZHO	11	13	18	–
SAR	–	16	14	–
DOO	–	–	–	17
BOR	–	–	–	18
ALO	R	3	5	11
ALB	14	R	11	9
BOT	R	18	17	–
GAS	8	9	R	R
TSU	R	11	15	R
STR	13	R	R	16
RIC	R	–	16	–
LAT	R	–	–	–
MSC	W	–	–	–

Tabela 31 – Posição de chegada de cada piloto por ano – Hungria.

Piloto	2022	2023	2024	2025
PIA	–	5	1	2
NOR	7	2	2	1
VER	1	1	5	9
PER	5	3	7	–
RUS	3	6	8	3
LEC	6	7	4	4
HAM	2	4	3	12
BOR	–	–	–	6
SAI	4	8	6	14
LAW	–	–	–	8
ALO	8	9	11	5
STR	11	10	10	7
VET	10	–	–	–
ANT	–	–	–	10
HAD	–	–	–	11
HUL	–	14	13	13
RIC	15	13	12	–
MSC	14	–	–	–
ALB	17	11	14	15
TSU	19	15	9	17
BOT	20	12	16	–
MAG	16	17	15	–
ZHO	13	16	19	–
SAR	–	18	17	–
LAT	18	–	–	–
COL	–	–	–	18
OCO	9	R	18	16
GAS	12	R	R	19
BEA	–	–	–	R

Tabela 32 – Posição de chegada de cada piloto por ano – Italia.

Piloto	2022	2023	2024	2025
VER	1	1	6	1
LEC	2	4	1	4
NOR	7	8	3	2
RUS	3	5	7	5
PER	6	2	8	–
HAM	5	6	5	6
SAI	4	3	4	11
PIA	–	12	2	3
ALB	–	7	9	7
BOR	–	–	–	8
DEV	9	–	–	–
ANT	–	–	–	9
HAD	–	–	–	10
MSC	12	–	–	–
BEA	–	–	–	12
LAW	–	11	–	14
BOT	13	10	16	–
SAR	–	13	–	–
GAS	8	15	15	16
ZHO	10	14	18	–
COL	–	–	12	17
MAG	16	18	10	–
LAT	15	–	–	–
OCO	11	R	14	15
STR	R	16	19	18
HUL	–	17	17	W
ALO	R	9	11	R
RIC	R	–	13	–
TSU	14	W	R	13
VET	R	–	–	–

Tabela 33 – Posição de chegada de cada piloto por ano – Estados Unidos.

Piloto	2022	2023	2024	2025
VER	1	1	3	1
NOR	6	2	4	2
PER	4	4	7	–
RUS	5	5	6	6
LEC	3	20	1	3
VET	8	–	–	–
HUL	–	11	8	8
BEA	–	–	–	9
TSU	10	8	14	7
SAR	–	10	–	–
LAW	–	–	9	11
MAG	9	14	11	–
GAS	14	6	12	19
ANT	–	–	–	13
ALB	13	9	16	14
COL	–	–	10	17
ZHO	12	13	19	–
MSC	15	–	–	–
RIC	16	15	–	–
HAD	–	–	–	16
LAT	17	–	–	–
BOR	–	–	–	18
HAM	2	19	R	4
ALO	7	R	13	10
STR	R	7	15	12
OCO	11	R	18	15
PIA	–	R	5	5
BOT	R	12	17	–
SAI	R	3	2	R

APÊNDICE C – Análise dos Resultados dos Modelos

Essa seção apresenta uma análise aprofundada dos resultados, tais como: coeficientes da regressão linear, importância das variáveis no XGBoost, valores SHAP e análise dos resíduos.

C.1 Coeficientes da Regressão Linear

Tabela 34 – Coeficientes da regressão linear (janela expansiva). Piloto, equipe e composto omitidas para melhor exibição.

Feature	Bahrein	Arabia Saudita	Hungria	Italia	Estados Unidos
LapTime_prev	1.4394	1.1365	0.9380	1.1758	1.3690
LapNumber	-0.1763	-0.1962	-0.1782	-0.1239	-0.1751
TyreLife	0.0967	0.0447	0.1053	0.0888	0.0655
Year	-0.0666	-0.0831	-0.0917	-0.1729	–
TrackTemp	0.1115	–	-0.0665	-0.0304	-0.2922
Humidity	0.0283	0.0317	–	0.0937	0.0133
Pressure	0.0113	-0.0108	-0.0433	-0.0390	0.0780
WindSpeed	-0.0153	0.0017	0.0179	-0.0009	-0.0022
TempDelta	0.0334	0.0476	-0.0109	0.0094	–

C.2 Importância das Variáveis

Tabela 35 – Importância das variáveis no XGBoost (janela expansiva). Piloto, equipe e composto omitidas para melhor exibição.

Feature	Bahrein	Arabia Saudita	Hungria	Italia	Estados Unidos
LapTime_prev	63.0631	56.5727	111.6854	181.5314	64.6636
LapNumber	0.4751	7.3132	10.6359	5.1375	0.7443
TyreLife	0.3263	0.7095	1.4054	1.5548	0.4993
Year	0.4173	12.4646	1.0135	234.1758	–
TrackTemp	0.7683	–	1.9212	5.7827	1.2102
Humidity	0.4646	4.0842	–	0.6123	0.5886
Pressure	0.4560	0.5094	11.8768	0.5043	0.6633
WindSpeed	0.3537	0.7743	0.7138	0.6311	0.4258
TempDelta	0.6407	0.7646	1.0074	1.8512	–

C.3 SHAP Values

Tabela 36 – Importância média absoluta dos valores SHAP do XGBoost no conjunto de teste (*holdout* sequencial), por circuito.

Feature	Bahrein	Arabia Saudita	Hungria	Italia	Estados Unidos
LapTime_prev	0.6483	0.9470	1.9647	2.3441	1.5159
LapNumber	0.0567	0.1255	0.0549	0.0601	0.0539
TyreLife	0.0402	0.0148	0.0356	0.0549	0.0233
Year	0.0005	0.0246	0.0107	0.2659	–
TrackTemp	0.0415	–	0.0816	0.0411	0.1088
Humidity	0.0090	0.0571	–	0.0002	0.0529
Pressure	0.0544	0.0077	0.1446	0.0002	0.0362
WindSpeed	0.0322	0.0169	0.0014	0.0062	0.0206
TempDelta	0.0203	0.0152	0.0067	0.0093	–

Tabela 37 – Importância média absoluta dos valores SHAP do modelo híbrido (componente LSTM) no conjunto de teste (*holdout* sequencial), por circuito.

Feature	Bahrein	Arabia Saudita	Hungria	Italia	Estados Unidos
LapTime_prev	0.001120	0.015345	0.008244	0.008525	0.006981
LapNumber	0.002969	0.006078	0.020844	0.007537	0.021031
TyreLife	0.004326	0.004689	0.019066	0.015899	0.004060
Year	0.000000	0.000000	0.000000	0.000000	–
TrackTemp	0.000985	–	0.000775	0.011736	0.001530
Humidity	0.005302	0.005034	–	0.006918	0.000000
Pressure	0.000001	0.002083	0.000170	0.000752	0.000572
WindSpeed	0.003826	0.013942	0.002830	0.014931	0.005071
TempDelta	0.001313	0.009410	0.001156	0.005399	–

C.4 Análise de Resíduos

Tabela 38 – Desvio padrão dos resíduos (s) no treino (validação) e no teste (*holdout* sequencial) por circuito – Regressao Linear (janela expansiva).

Circuito	s_{treino}	s_{teste}
Bahrein	0.3059	0.3140
Arabia Saudita	0.3406	0.3185
Hungria	0.3238	0.3412
Italia	0.2930	0.2715
Estados Unidos	0.3518	0.3330

Tabela 39 – Desvio padrão dos resíduos (s) no treino (validação) e no teste (*holdout* sequencial) por circuito – XGBoost (janela expansiva).

Circuito	s_{treino}	s_{teste}
Bahrein	0.2821	0.3122
Arabia Saudita	0.3168	0.3175
Hungria	0.3201	0.5466
Italia	0.3309	0.3413
Estados Unidos	0.3117	0.3378

Tabela 40 – Desvio padrão dos resíduos (s) no treino (validação) e no teste (holdout sequencial) por circuito – Modelo Híbrido.

Circuito	s_{treino}	s_{teste}
Bahrein	0.2521	0.3164
Arabia Saudita	0.3088	0.3186
Hungria	0.3110	0.3462
Italia	0.2675	0.2657
Estados Unidos	0.3186	0.3301