UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Estimação de Altitude durante Curvas de Reversão em Aeronaves Agrícolas com Dados de *Differential Global Positioning System*

Felipe Caldoncelli Barra

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

Felipe Caldoncelli Barra

Estimação de Altitude durante Curvas de Reversão em Aeronaves Agrícolas com Dados de *Differential Global Positioning System*

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Afonso Paiva Neto

Versão original

São Carlos

2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Caldoncelli Barra, Felipe
C145e Estimação de Altitude durante Curvas de Reversão em Aeronaves Agrícolas com Dados de Differential Global Positioning System / Felipe Caldoncelli Barra; orientador Afonso Paiva. -- São Carlos, 2026.
79 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2026.

1. Aviação Agrícola. 2. Curva de reversão. 3. Investigação de ocorrências aeronáuticas. 4. Estimação de altitude. 5. Differential Global Positioning System. I. Paiva, Afonso, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Felipe Caldoncelli Barra

Estimação de Altitude durante Curvas de Reversão em Aeronaves Agrícolas com Dados de *Differential Global Positioning System*

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Afonso Paiva

Original version

São Carlos

2026

À memória dos pilotos agrícolas que, no exercício de sua missão, não lograram regressar da lavoura ao final da jornada: a eles dedico este trabalho, em respeito à sua coragem e ao impacto de sua ausência.

AGRADECIMENTOS

Agradeço ao meu pai, cuja conduta e ponderação sempre se constituíram em exemplo e referência que orientaram minha trajetória. Em momentos de maior fragilidade, como este, compreendo com ainda mais clareza que a verdadeira força reside na mente e na forma como escolhemos conduzir nossos passos.

Agradeço à minha mãe, por me revelar, com simplicidade e firmeza, o sentido do amor e do dever.

Agradeço à minha esposa, pelo companheirismo, pela compreensão e pelo apoio constante.

Agradeço à minha filha, por me fazer ver mais longe e ressignificar prioridades.

Agradeço ao Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA), por propiciar os meios, pela confiança e por acreditar neste trabalho.

"Eu olho para um carro como olho para qualquer outra coisa."
Warren Buffett

Per aspera ad astra

RESUMO

CALDONCELLI BARRA, F. **Estimação de Altitude durante Curvas de Reversão em Aeronaves Agrícolas com Dados de *Differential Global Positioning System***. 2026. 79 p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2026.

Este trabalho investiga a viabilidade de estimar a variação de altitude durante curvas de reversão em operações aeroagrícolas a partir de registros do sistema AgNav–Guia, visando apoiar a reconstrução do perfil vertical em cenários nos quais a altitude GPS não está disponível. Os dados foram coletados, tratados e segmentados por manobra, e modelos de regressão supervisionada foram avaliados para a predição de Δalt, com análise de desempenho por métricas globais e por curva e inspeção dos piores casos. No escopo experimental adotado, o CatBoost Regressor apresentou desempenho superior entre os modelos testados, com erro médio absoluto (MAE) relativamente pequeno quando comparados à escala típica da variável-alvo. Observou-se comportamento consistente no regime predominante, embora com presença de cauda longa de erro associada a um subconjunto de curvas com maior dificuldade preditiva. A análise das curvas com MAE por curva acima do percentil 95 sugere associação dos maiores erros a combinações específicas de condições operacionais e a possíveis fatores não observados nos registros disponíveis, o que pode limitar a generalização em situações particulares. Conclui-se que a abordagem é promissora para apoiar a reconstrução do perfil vertical de manobras e pode contribuir para o aprimoramento de análises técnicas e investigações de ocorrências, com potencial de auxiliar a prevenção de novos eventos por meio de melhor compreensão de condições operacionais e fatores contribuintes.

Palavras-chave: Aviação Agrícola. Curva de reversão. Investigação de ocorrências aeronáuticas. Estimação de altitude. *Differential Global Positioning System*

ABSTRACT

CALDONCELLI BARRA, F. **Altitude Estimation during Reversal Turns in Agricultural Aircraft Using *Differential Global Positioning System Data***. 2026. 79 p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2026.

This study investigates the feasibility of estimating altitude variation during reversal turns in agricultural aviation operations using records from the AgNav–Guia system, in order to support vertical-profile reconstruction when GPS altitude is not available. Flight data were collected, preprocessed, and segmented by maneuver, and supervised regression models were evaluated to predict Δalt, including performance assessment through global and per-turn metrics and an inspection of the worst cases. Within the adopted experimental scope, the CatBoost Regressor outperformed the other tested models, achieving a mean absolute error (MAE) that is relatively small when compared to the typical scale of the target variable. The model showed consistent behavior in the predominant operating regime, although a long error tail was observed, associated with a subset of turns with higher predictive difficulty. The analysis of turns with per-turn MAE above the 95th percentile suggests that larger errors are related to specific combinations of operating conditions and to potential unobserved factors in the available records, which may limit generalization in particular situations. Overall, the proposed approach is promising for supporting maneuver vertical-profile reconstruction and may contribute to improved technical analyses and occurrence investigations, with potential to help prevent future events by enhancing the understanding of operational conditions and contributing factors.

Keywords: Agricultural aviation. Reversal turn. Accident investigation. Altitude estimation. *Differential Global Positioning System.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANAC	Agência Nacional de Aviação Civil
CENIPA	Centro de Investigação e Prevenção de Acidentes Aeronáuticos
CNA	Confederação da Agricultura e Pecuária do Brasil
CPU	*Central Processing Unit*
DGPS	GPS Diferencial
EFB	*Exclusive Feature Bundling*
ET	*Extra Trees*
EUA	Estados Unidos da América
FAA	*Federal Aviation Administration*
FDM	*Flight Data Monitoring*
GOSS	*Gradient-based One-Side Sampling*
GPU	*Graphics Processing Unit*
GPS	*Global Positioning System*
GS	Velocidade sobre o solo
LGBM	LightGBM
LSTM	*Long Short-Term Memory*
MAE	Erro Absoluto Médio
MAPA	Ministério da Agricultura e Pecuária
MSE	Erro Quadrático Médio
OLS	*Ordinary Least Squares*
P95	Percentil 95
PIB	Produto Interno Bruto
RMSE	Raiz do Erro Quadrático Médio
SHAP	*SHapley Additive exPlanations*

SINDAG	Sindicato Nacional das Empresas de Aviação Agrícola
SIPAER	Sistema de Investigação e Prevenção de Acidentes Aeronáuticos
TT	Tempo de treinamento
USPSC	Campus USP de São Carlos
UTC	Tempo Universal Coordenado

LISTA DE SÍMBOLOS

X	Vetor (ou conjunto) de variáveis explicativas (preditoras)
X_j	j-ésima variável explicativa
Y	Variável resposta (alvo)
y_i	Valor observado da variável resposta na i-ésima observação
$\hat{y}_i$	Valor predito da variável resposta na i-ésima observação
$\mathbb{E}$	Operador de esperança (valor esperado). No trabalho, $\mathbb{E}[\hat{y}]$ denota o valor esperado das predições
β_0	Termo constante (intercepto) em modelos lineares
β_j	Coeficiente associado à j-ésima variável explicativa em modelos lineares
n	Número de observações (amostras)
p	Número de variáveis explicativas (preditoras)
λ	Parâmetro de regularização (penalização)
$RSS(\lambda)$	Soma de quadrados dos resíduos com penalização (função de custo do Ridge)
ϕ_j	Valor SHAP associado à variável j (contribuição de j para a predição)
|SHAP|	Magnitude absoluta do valor SHAP
MAE	Erro absoluto médio
$P95$(MAE)	Percentil 95 do erro absoluto médio
R^2	Coeficiente de determinação
ψ	Proa (ângulo de rumo/heading)
ψ_{in}	Proa no início da curva
$\Delta\psi_{in}$	Variação angular em relação ao início da curva
Δalt	Variação de altitude (diferença de altitude)

SUMÁRIO

1 INTRODUÇÃO

O setor do agronegócio desempenha um papel fundamental na economia brasileira, sendo responsável, segundo a Confedereção da Agricultura e Pecuária do Brasil (CNA, 2025) por um quarto do Produto Interno Bruto (PIB) do país no ano de 2024. No mesmo período a soma de bens e serviços gerados no agronegócio chegou a R$ 2,72 trilhões, tendo a maior parcela deste valor, 70%, sendo oriunda do ramo agrícola (R$ 1,9 trilhão).

Neste contexto, segundo o Ministério da Agricultura e Pecuária (MAPA, 2025), a aviação agrícola busca proteger as lavouras e auxiliar o desenvolvimento da agricultura através da realização de serviços como semeadura, aplicação de fertilizantes e agrotóxicos (químicos e biológicos) para proteção das lavouras, além do combate a incêndios em todos os tipos de vegetação. A aviação agrícola é empregada em mais de 15 culturas no Brasil, entre estas: cana de açúcar, café, arroz, soja, milho, trigo, algodão, banana, laranja, etc.

Segundo Gomes (2024), estavam registradas na Agência Nacional de Aviação Civil (ANAC), no ano de 2024, 2.772 aeronaves para uso agrícola no Brasil. Dentre estas, 99% eram compreendidas por aviões e apenas 1% por helicópteros. Desta forma, a frota de aeronaves para uso agrícola, seguindo os critérios empregados pelo autor, representava 12,4% da frota nacional.

Muito embora o primeiro voo agrícola no Brasil tenha ocorrido em 1947, foi apenas nos final dos anos 90 que diversas inovações tecnológicas como novos tipos de pontas de pulverização, barras aerodinâmicas e aprimoramentos nos equipamentos nacionais permitiram a expansão deste segmento. Mas o maior destes avanços, segundo o Sindicato Nacional das Empresas de Aviação Agrícola (SINDAG, 2025), foi o GPS (*Global Positioning System*).

Para os pilotos, sistemas de navegação GPS auxiliam na execução da aplicação de insumos agrícolas, guiando-os com precisão para as áreas designadas. Para o contratante, os relatórios de aplicação gerados por estes sistemas são utilizados como certificação da execução e qualidade do serviço. Mas, muito embora não tenham sido desenvolvidos para esta finalidade, vêm sendo utilizados para a investigação de acidentes aeronáuticos, realizados pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA). Por meio dos dados registrados é possível recuperar o histórico de voos, a trajetória da aeronave, bem como determinar parâmetros operacionais empregados (CENIPA,2025).

Desde 2015 foram registrados 451 acidentes na aviação agrícola pelo Sistema de Investigação e Prevenção de Acidentes Aeoronáuticos (SIPAER, 2025), representando 27,4% do total de acidentes aéreos ocorridos no Brasil. Entre os tipos de ocorrência da aviação agrícola, 21% foram classificadas como ‘Perda de controle em voo’.

Nas operações aeroagrícolas, são realizadas muitas curvas para reposicionamento para pulverização, conhecidas como curvas de reversão, manobra que expõe a aeronave a uma ampla variação de parâmetros, consequentemente, deixando-a mais propensa à perda de controle em voo. A informação de altitude, nestes casos, é um fator importante para determinação da trajetória de voo.

Entretanto, os sistemas de navegação possuem diferentes características de gravação de dados e, em alguns modelos, as informações de altitude não são registradas, impossibilitando a determinação da trajetória no eixo vertical.

1.1 Motivação

A análise detalhada das curvas de reversão contribui para compreensão do comportamento dinâmico das aeronaves agrícolas e para a identificação de condições de risco, como a aproximação da velocidade de estol. Contudo, alguns sistemas de navegação empregados em aeronaves agrícolas não registram a altitude GPS, limitando a capacidade de calcular outros parâmetros que dependem diretamente dessa informação. Sem esses dados, a investigação de acidentes aeronáuticos perde elementos importantes para a análise.

Nesse contexto, a utilização de técnicas de aprendizagem de máquina, em especial o emprego de modelos de regressão, pode suprir a ausência do registro direto de altitude. O modelo, desenvolvido a partir de dados de equipamentos que registram a altitude GPS, poderia então ser aplicado naqueles que não possuem esta mesma capacidade. Como consequência, seria possível estimar a elevação do avião durante as curvas de reversão, mesmo nos equipamentos mais simples.

1.2 Objetivo

Este trabalho tem como objetivo desenvolver e avaliar modelos de regressão capazes de estimar a altura de aeronaves agrícolas durante as curvas de reversão, em sistemas de navegação que não registram diretamente esse parâmetro. Para isso, serão exploradas variáveis disponíveis — como velocidade sobre o solo, proa verdadeira e funcionamento do sistema de pulverização, oriundas de equipamentos do tipo ‘AgNav – Guia’ — para treinar modelos preditivos que estimem com a maior precisão possível a altura do avião durante as curvas de reversão. Como consequência, estes modelos desenvolvidos poderiam ser utilizados para estimação de altura em aeronaves equipadas com sistemas de navegação incapazes de registrar este parâmetro.

1.3 Organização do texto

Este trabalho de conclusão de curso é organizado da seguinte forma. No Capítulo 2, é apresentada uma fundamentação teórica e uma revisão bibliográfica, a fim de possibilitar ao leitor compreender conceitos de aviação e os estudos realizados na área relacionados ao escopo deste trabalho. O Capítulo 3 introduz os conceitos fundamentais de ciência de dados e os principais modelos de regressão utilizados. No Capítulo 4, é detalhada a metodologia adotada, incluindo a coleta de dados e as ferramentas empregadas. Em seguida, no Capítulo 5, são apresentados os resultados obtidos, seguidos das conclusões no Capítulo 6.

2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta os conceitos fundamentais e os trabalhos anteriores relacionados à segurança de voo que fundamentam o estudo. O leitor aprenderá sobre a análise de dados de voo em operações aeroagrícolas e a importância do monitoramento operacional e da caracterização dos parâmetros de voo para a identificação de riscos e a prevenção de ocorrências. Também são revisadas abordagens da literatura baseadas em registros de voo, situando o presente trabalho no contexto da segurança operacional.

2.1 Fundamentação Teórica

O objetivo da fundamentação teórica é apresentar conceitos aeronáuticos fundamentais ao entendimento do problema de pesquisa. A revisão bibliográfica deste trabalho visa compreender o estado atual de pesquisas referentes ao tema de estimação de parâmetros de voo a partir de sistemas de navegação agrícolas e/ou à estimação de variações de altitude de uma aeronave em função da variação da sua velocidade.

2.1.1 Equipamentos DGPS

A evolução da tecnologia de posicionamento global tornou acessível à aviação agrícola um sistema de maior precisão, o Differential Global Positioning System, ou GPS Diferencial em português. Essa tecnologia reduziu as perdas por sobreposição de áreas, por exemplo.

O GPS Diferencial (DGPS) emprega um receptor fixo (“Estação de Referência”), em coordenadas conhecidas, que recebe os mesmos sinais dos satélites que o receptor móvel do usuário. Ao comparar os dados recebidos com sua posição real, calcula os erros de cada satélite e transmite essas correções — geralmente por rádio — ao receptor diferencial do usuário. Este, por sua vez, processa simultaneamente o sinal GPS bruto (com erros de 10–15 m) e o sinal diferencial (contendo os ajustes de erro), gerando um posicionamento depurado que reduz a incerteza para cerca de 1–5 m e, em configurações otimizadas, até 30 cm. A qualidade da correção é monitorada pela “idade do sinal diferencial”, isto é, o tempo decorrido (em segundos) desde a última atualização das correções, sendo que valores menores asseguram maior precisão (Araújo, 2005).

Os dados registrados então, geralmente bastante precisos, tornam-se ferramentas valiosas para a análise dos perfis de voo. Esses dados são registrados de formas diferentes de acordo com o fabricante, podendo variar desde arquivos de dados geoespaciais (`.kmz`), contendo apenas pontos geográficos adquiridos à razão de tempo constante, até outros formatos proprietários que podem ser decodificados por meio de softwares específicos,

gerando planilhas tabulares com os parâmetros registrados.

2.1.2 Conceitos de aviação

Altitude GPS É a altura da aeronave em relação a um modelo elipsoidal do planeta Terra (Mississipi, 2020).

Ângulo de trajetória de voo É o ângulo vertical entre o vetor da trajetória de voo e o horizonte. (APS, 2025).

Curva de reversão Manobra destinada a inverter o sentido de aplicação de insumos, geralmente compreendendo uma variação de 180º de proa.

Estol Redução da sustentação causada pela separação do fluxo de ar da superfície da asa, provocada pela ultrapassagem do ângulo de ataque crítico (FAA, pág 5-25, 2024).

Proa verdadeira Direção em relação ao norte verdadeiro, medida em graus no sentido horário. Aplicando-se a declinação magnética obtém-se a proa magnética (FAA, pág 5-25, 2024).

Razão de curva Quantidade de graus que a aeronave modifica sua proa por unidade de tempo, geralmente representada em graus/segundo (FAA, 2024).

Velocidade sobre o solo Velocidade real da aeronave em relação ao solo, variando em função do vento (FAA, pág 8-9, 2024).

2.1.3 Dados de voo

Os dados de voo são a base para que sejam realizadas análises de tendência ou, como é o objetivo deste trabalho, o desenvolvimento de modelos de predição. No contexto da aviação, cada série de dados que representa uma variável do comportamento da aeronave ou de um de seus sistemas é conhecida como parâmetro.

Os parâmetros de voo podem ser analisados em comparação a um comportamento esperado ou utilizados para, por meio de um processo de derivação, determinar outros parâmetros que não são adquiridos por sensores ou registrados em memória.

2.1.4 Dados de voo na investigação de acidentes agrícolas

A importância da utilização de dados de voo para investigação pode ser exemplificada observando-se o Relatório Final de acidente envolvendo uma aeronave AT-502A, ocorrido em 18 de novembro de 2022 (CENIPA, 2024b), que estava equipada com um sistema de navegação ‘AgNav-Guia’.

Os dados deste equipamento são decodificados por meio de software (*NavView*), gerando uma planilha tabular cujos principais campos para investigação são: horário GPS

de cada registro, coordenadas geográficas, velocidade sobre o solo, altitude GPS e condição de funcionamento do sistema de pulverização (ligado/desligado).

Com base nesses dados, juntamente com conhecimentos de engenharia aeronáutica e técnicas de investigação de acidentes, foi possível criar um modelo que calculava o ângulo de trajetória de voo, a razão vertical, a velocidade linear da aeronave e estimava a inclinação de asas para voo nivelado e para o topo das curvas de reversão.

Durante as curvas de reversão, as aeronaves ganham altura e iniciam uma curva para o sentido oposto. Como consequência, ocorre a diminuição de velocidade e o aumento da inclinação de asas. À medida que a aeronave inclina, a velocidade de estol aumenta. Caso não seja mantida uma margem segura entre a velocidade da aeronave e a velocidade de estol, pode haver propensão à perda de controle.

Figura 1 – Aeronave agrícola iniciando uma curva de reversão.

Fonte: Curbside Classic. Disponível em: https://www.curbsideclassic.com/rampside-classic/rampside-fieldside-classic-1967-aero-commander-s-2d/. Acesso em: dez. 2025.

Na investigação citada, evidenciou-se que, nas curvas realizadas no setor do acidente, a velocidade no topo aproximava-se à de estol. O emprego de subidas acentuadas, no intuito de dinamizar o circuito de aplicação, contribuía para a redução das margens de velocidade (CENIPA, 2024b). A compreensão da ocorrência só foi alcançada devido aos parâmetros disponíveis na memória do equipamento, inclusive a altitude GPS, base para obtenção de parâmetros não diretamente registrados. Sua utilização em investigações contribui para evitar ocorrências similares, reduzindo fatalidades e perdas materiais.

2.2 Revisão Bibliográfica

A utilização de técnicas de aprendizado de máquina no monitoramento de dados de voo está em crescente aumento. Segundo Jasra et al. (2018), essa abordagem permite extrair informações operacionais relevantes, com foco na detecção de anomalias e seleção de variáveis importantes em programas de monitoramento de dados de voo (Flight Data Monitoring – FDM).

No campo da investigação de ocorrências, a quantidade de trabalhos é limitada. Mahgortey et al. (2020) propõe metodologia para selecionar parâmetros mais relevantes de dados de voo, incluindo filtragem por correlação, exigências regulatórias, clusterização e análise de variância, aplicada à fase de aproximação de uma aeronave.

Se para aeronaves de grande porte existem muitos parâmetros disponíveis, na aviação agrícola a situação é oposta, pois, o normal é a escassez de dados. Daí surge a necessidade de estimar parâmetros adicionais a partir dos poucos registrados.

Existem poucos estudos voltados à análise de dados de voo em operações agrícolas, visto que sua utilização costuma a restringir-se a grupos de investigação de acidentes aeronáuticos. Ademais, não foi identificado nenhum estudo que utilize aprendizado de máquina para estimar altitudes, ou outro parâmetro, em curvas de reversão.

Esse vácuo na literatura justifica o presente trabalho, que se propõe a aplicar modelos de aprendizado de máquina para estimar a altitude da aeronave, contribuindo para a investigação de acidentes aeronáuticos.

Foram identificadas três publicações consideradas mais similares aos objetivos deste trabalho. A Tabela 1 apresenta um extrato de suas principais características relativas ao tratamento dos dados analisados e ao desenvolvimento de modelos de aprendizado de máquina.

Mississipi (2020), realizou um estudo com cerca de 30.000 registros de voos agrícolas nos Estados Unidos da América. Seu objetivo era identificar o comportamento dos voos, a fim de integrá-los a sistemas aéreos não tripulados em um mesmo espaço aéreo. Neste estudo, foram identificados alguns parâmetros mais importantes para a análise, sendo estes: ângulos de trajetória de voo antes e após a pulverização, velocidade e altura durante as pulverizações e em cruzeiro. Como resultado, identificaram-se comportamentos gerais adotados pelas aeronaves durante a operação aeroagrícola.

As duas outras publicações encontradas tratam-se de Relatórios Finais de Investigação de Acidentes Aeronáuticos, ou seja, os resultados de uma investigação. Consequentemente, abordam, cada qual, casos específicos para determinada aeronave e condições existentes. Muito embora semelhantes no método e objetivos, divergem quanto à fonte e disponibilidade de dados. Enquanto CENIPA (2024b) tinha à disposição dados de

um sistema ‘AgNav-Guia’, capaz de registrar planilhas tabulares com dados completos como altura e tempo, CENIPA (2024a) dispunha apenas de um arquivo ‘.kml’ contendo coordenadas geográficas.

Neste último caso, mesmo parâmetros básicos de voo, como a velocidade sobre o solo, não são registrados. Tal situação foi observada por CENIPA (2024a) durante a investigação do acidente com a aeronave A188B, onde os dados disponíveis eram oriundos de arquivo ‘.kml’, adquirido a 5 Hz. Como solução, a velocidade sobre o solo foi obtida por meio de derivação matemática da distância entre pontos em função do tempo.

Outro aspecto importante para os objetivos deste estudo é a identificação das curvas de reversão. A mesma situação foi abordada por Mississipi (2020), que utilizou a variação angular entre o último ponto da pulverização anterior e o primeiro da subsequente na identificação desses segmentos.

No tocante à razão de curva, Mississipi (2020), calculou este parâmetro para detectar tendências de operação. Enquanto isso, em ambas as investigações, este dado foi utilizado como um dos valores para estimar a inclinação das asas no topo das curvas de reversão.

Por fim, Mississipi (2020) recorda que existe diferença de performance entre diversos modelos de aeronaves agrícolas, mais notadamente em relação à velocidade. Segundo o estudo, o conhecimento da performance nominal poderia auxiliar na validação dos dados e na classificação das aeronaves.

Tabela 1 – Comparação de estudos em análise de dados de voo

Autor	**Dados**	**Análise de Dados**	**Machine Learning**
Mississippi (2020)	30 000 registros de voo (EUA)	Sim	Não
CENIPA (2024b)	AT-502A (DGPS *AgNav-Guia*)	Sim	Não
CENIPA (2024a)	A188B (*.kml*)	Sim	Não

Fonte: Autor.

3 CONCEITOS FUNDAMENTAIS EM CIÊNCIA DE DADOS

A Ciência de Dados reúne técnicas estatísticas e computacionais que permitem extrair informações relevantes e construir modelos capazes de representar fenômenos complexos. No contexto deste trabalho, esses conceitos são aplicados à estimação da altura de aeronaves agrícolas durante curvas de reversão, utilizando outras variáveis registradas por sistemas de navegação.

A análise de dados de voo, composta por variáveis contínuas e inter-relacionadas — como velocidade sobre o solo, coordenadas geográficas e seus parâmetros derivados — requer o uso de modelos de regressão e aprendizado de máquina capazes de capturar padrões e dependências não lineares. Esses métodos permitem transformar medições complexas em estimativas, oferecendo uma alternativa viável para a reconstrução do parâmetro de altura em aeronaves que não o registram diretamente.

3.1 Aprendizado de máquina

O aprendizado de máquina (*machine learning*) tem como objetivo desenvolver métodos que permitam aos sistemas computacionais aprender e aprimorar seu desempenho por meio da experiência, sem a necessidade de programação explícita. Segundo Lindholm *et al.* (2022), o aprendizado de máquina pode ser entendido como um processo de "programação por exemplo", no qual algoritmos identificam padrões e relações em conjuntos de dados e os utilizam para realizar previsões ou tomar decisões.

Essa abordagem representa uma mudança de paradigma em relação aos métodos tradicionais, pois transfere para os dados a responsabilidade de guiar a criação do modelo, em vez de depender exclusivamente de regras definidas por programadores. Ainda conforme Lindholm *et al.* (2022), qualquer sistema de aprendizado de máquina baseia-se em três componentes fundamentais: os dados, o modelo matemático e o algoritmo de aprendizado. O modelo constitui uma representação formal e simplificada das relações existentes entre variáveis, enquanto o algoritmo tem como função ajustar os parâmetros desse modelo a partir dos dados observados, buscando minimizar o erro e maximizar a capacidade preditiva.

Esse processo iterativo de ajuste permite que o sistema generalize o conhecimento adquirido e o aplique a novos conjuntos de dados, o que justifica sua ampla aplicabilidade em diferentes áreas. O aprendizado de máquina pode ser classificado, de forma geral, em supervisionado e não supervisionado, conforme o tipo de dado utilizado e o objetivo da análise.

No aprendizado supervisionado, o modelo é treinado com exemplos rotulados — pares de entrada e saída conhecidos — possibilitando a realização de tarefas como regressão e classificação. Por outro lado, o aprendizado não supervisionado trabalha com dados não rotulados, buscando identificar padrões, agrupamentos ou representações reduzidas das variáveis observadas, como ocorre nas técnicas de *clustering* e redução de dimensionalidade. Esses dois paradigmas representam abordagens complementares para a extração de conhecimento a partir de dados (Lindholm *et al.*, 2022).

3.2 Modelos de Regressão

Os modelos de regressão constituem a base da análise preditiva na ciência de dados, permitindo estimar relações funcionais entre uma variável dependente e um conjunto de variáveis independentes, de modo a compreender como alterações nas variáveis explicativas influenciam a resposta observada.

Segundo Harrell (2015), o propósito da regressão é construir uma função que represente os mecanismos subjacentes aos dados, seja em contextos descritivos ou preditivos. Em sua forma clássica, os modelos lineares — como a regressão linear simples e suas extensões regularizadas, como a Regressão Ridge — assumem uma relação linear entre os preditores e a variável de resposta, oferecendo boa interpretabilidade e estabilidade estatística, sobretudo quando o número de variáveis é limitado e as correlações são moderadas.

No entanto, em situações em que as relações entre as variáveis são complexas e não lineares, torna-se necessário recorrer a modelos baseados em aprendizado de máquina, capazes de capturar padrões mais sofisticados. Modelos como Extra Trees (ET), Random Forest (RF), CatBoost e LightGBM (LGBM), de acordo com Lindholm *et al.* (2022), empregam estruturas de árvores de decisão e métodos de *ensemble* para explorar interações e não linearidades entre as variáveis, resultando em alto poder preditivo e capacidade de generalização.

Regressão Linear

Um modelo de Regressão Linear (LR) assume que a função de regressão $E(Y \mid X)$ é linear em relação às variáveis independentes $X_1, X_2, \ldots, X_p$. De acordo com Hastie, Tibshirani e Friedman (2009), apesar de sua formulação ter origem na era pré-computacional, os modelos lineares permanecem amplamente utilizados por sua simplicidade, interpretabilidade e desempenho competitivo em cenários de dados limitados ou baixa variabilidade. O modelo pode ser expresso pela Equação 3.1:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{3.1}$$

onde β_0 representa o intercepto e β_j os coeficientes que descrevem o impacto de cada variável X_j sobre a resposta Y. Esses parâmetros são geralmente estimados pelo método dos mínimos quadrados ordinários (OLS), que busca minimizar a soma dos quadrados dos resíduos. Conforme Harrell (2015), a regressão linear fornece uma estrutura analítica sólida e interpretável, servindo como referência para a validação e comparação de modelos mais complexos.

Regressão Ridge

A Regressão Ridge é uma extensão dos modelos lineares clássicos que incorpora um termo de penalização aos coeficientes, reduzindo sua magnitude para mitigar efeitos de multicolinearidade e sobreajuste. O modelo busca minimizar a seguinte função de custo (Hastie; Tibshirani; Friedman, 2009):

$$RSS(\lambda) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.2}$$

onde λ é o hiperparâmetro de regularização que controla a penalidade imposta aos coeficientes. Esse método oferece um compromisso entre viés e variância, melhorando a estabilidade e a generalização do modelo, especialmente quando as variáveis preditoras apresentam alta correlação (Lindholm *et al.*, 2022).

Extra Trees (ET)

O Extra Trees (*Extremely Randomized Trees*) é um método de *ensemble* baseado em árvores de decisão que introduz elevado grau de aleatoriedade na construção das árvores, tanto na seleção das variáveis quanto na escolha dos pontos de divisão. Conforme Geurts, Ernst e Wehenkel (2006), essa estratégia reduz a correlação entre as árvores e melhora a capacidade de generalização do modelo, mesmo com pequeno aumento do viés. Diferentemente do Random Forest, o Extra Trees utiliza o conjunto completo de dados, evitando a reamostragem por *bootstrap*, e define aleatoriamente os pontos de corte para cada divisão, avaliando o melhor entre um subconjunto de divisões possíveis. Essa dupla randomização aumenta a diversidade entre os estimadores individuais e, quando combinada pela média das previsões, reduz a variância do modelo final.

Random Forest (RF)

O Random Forest (RF) é um método de *ensemble* baseado em árvores de decisão, que aprimora a técnica de *bagging* por meio da introdução de aleatoriedade no processo de construção das árvores. De acordo com Hastie, Tibshirani e Friedman (2009), o modelo consiste em gerar múltiplas árvores a partir de subconjuntos de dados obtidos por reamostragem (*bootstrap*) e, a cada divisão, selecionar aleatoriamente um subconjunto das variáveis preditoras para determinar o ponto de corte mais adequado. Essa estratégia reduz a correlação entre as árvores e, consequentemente, a variância do modelo final, mantendo

baixo o viés e evitando o sobreajuste. As predições do conjunto são obtidas pela média (em regressão) ou pela votação majoritária (em classificação), conferindo ao modelo robustez e boa precisão preditiva.

CatBoost

O CatBoost é um algoritmo de *gradient boosting* sobre árvores de decisão, projetado para reduzir o viés estatístico presente em implementações tradicionais desse método. Conforme Prokhorenkova *et al.* (2018), o CatBoost introduz dois avanços principais: o *ordered boosting* e o tratamento ordenado de variáveis categóricas, ambos concebidos para eliminar o *prediction shift* (uma forma de vazamento de informação). Mesmo em bases compostas exclusivamente por variáveis numéricas, o mecanismo de *ordered boosting* corrige o desvio de predição por meio de permutações aleatórias dos dados de treino, garantindo que a estimativa dos resíduos seja feita de forma imparcial.

LightGBM (LGBM)

O LightGBM (*Light Gradient Boosting Machine*) é uma implementação eficiente de *gradient boosting* sobre árvores de decisão que introduz inovações estruturais para acelerar o treinamento e reduzir o consumo de memória. Conforme Ke *et al.* (2017), o LightGBM adota duas técnicas centrais: *Gradient-based One-Side Sampling* (GOSS) — que reduz a amostra focando em observações com gradientes maiores — e *Exclusive Feature Bundling* (EFB) — que combina características mutuamente exclusivas para reduzir a dimensionalidade sem perda relevante de informação. Além disso, ao contrário do crescimento nível a nível (*level-wise*), o LightGBM cresce árvores folha a folha (*leaf-wise*), selecionando a folha com maior redução de perda para dividir, o que pode melhorar a convergência com complexidade controlada.

4 METODOLOGIA

Este capítulo descreve o processo metodológico adotado no desenvolvimento do estudo, detalhando as etapas desde a obtenção e organização dos dados até a construção do conjunto analítico utilizado na modelagem. O leitor aprenderá sobre os procedimentos de pré-processamento, segmentação e tratamento dos registros de voo, bem como sobre a definição e transformação das variáveis relevantes. Destaca-se o processo de *feature engineering*, no qual parâmetros derivados são construídos a partir dos dados brutos com o objetivo de representar de forma mais adequada a dinâmica operacional das curvas analisadas. A metodologia apresentada busca assegurar coerência entre os dados, as variáveis construídas e os objetivos do estudo, fornecendo uma base consistente para as etapas de modelagem e avaliação apresentadas posteriormente.

4.1 Estrutura geral

Este trabalho tem como objetivo estimar a altitude de aeronaves agrícolas durante curvas de reversão, por meio da aplicação de técnicas de aprendizado de máquina a dados de voo obtidos por sistemas de navegação DGPS. A abordagem proposta é de natureza quantitativa, fundamentada em métodos de análise de dados e regressão supervisionada, com o propósito de identificar relações entre variáveis de desempenho da aeronave e sua variação de altitude (Δalt).

Foram reunidos registros de voo provenientes de sistemas embarcados em aeronaves agrícolas, contendo parâmetros operacionais — em especial a altitude GPS — e informações geográficas. A partir desses registros, selecionaram-se e derivaram-se variáveis relevantes ao problema de estimação (*feature engineering*), permitindo estruturar uma base de dados adequada para o treinamento dos modelos.

Em seguida, as curvas de reversão foram identificadas e extraídas dos registros, de modo a compor o conjunto de dados de interesse para a modelagem. Sobre esse conjunto segmentado, aplicaram-se diferentes modelos de regressão, comparando-se seus desempenhos na previsão do Δalt a partir das variáveis selecionadas.

Por fim, ao modelo de melhor desempenho, aplicaram-se os dados de teste, sendo gerada uma visualização tridimensional (*3D*) que permitiu reconstruir o perfil vertical das manobras e avaliar a viabilidade prática da técnica proposta.

Nas subseções seguintes, são detalhados os procedimentos empregados em cada uma dessas etapas, descrevendo-se as fontes de dados, o processo de preparação e construção das variáveis, os critérios de identificação das curvas de reversão, os modelos utilizados e

os métodos de avaliação de desempenho.

4.2 Coleta e preparação dos dados

Os dados utilizados neste estudo foram provenientes de dispositivos de navegação do tipo AgNav–Guia, amplamente empregados em aeronaves agrícolas brasileiras. Os registros analisados integraram a base de dados do Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA) e corresponderam a operações aeroagrícolas normais, não necessariamente associadas a ocorrências aeronáuticas. Para fins deste trabalho, a utilização dos dados foi autorizada de forma descaracterizada, sem qualquer vínculo identificável com operadores ou locais específicos.

Para esta análise, foram utilizados dados de um único modelo de aeronave, o Air Tractor 502A, pertencentes a um único operador. Os registros abrangeram um período superior a dois anos, totalizando 1.278 arquivos de dados de voo. Dentre esses, foram selecionados 58 arquivos para conversão e análise detalhada.

Os arquivos originais, gerados em formato proprietário, foram convertidos para planilhas no formato `.csv` por meio do software *NavView*, disponibilizado pela própria fabricante AGNAV. Essa conversão permitiu a manipulação dos registros em ambiente analítico, preservando a estrutura temporal e espacial das observações originais.

A Tabela 2 apresenta as principais variáveis registradas pelos equipamentos AgNav–Guia, disponíveis nos dados de voo e aplicáveis ao problema de pesquisa.

Tabela 2 – Parâmetros utilizados no estudo

Parâmetro	**Descrição**
Horário GPS	Registro temporal de cada amostra (UTC).
Latitude	Coordenada geográfica em graus (°) e frações.
Longitude	Coordenada geográfica em graus (°) e frações.
Velocidade sobre o solo (GS)	Velocidade real da aeronave em relação ao solo (*mph*).
Proa verdadeira	Direção da trajetória da aeronave em relação ao norte verdadeiro (°).
Altitude GPS	Altura da aeronave em relação ao modelo elipsoidal da Terra (*ft*).
Spray	Estado de funcionamento do sistema de pulverização (On/Off).

Fonte: Autor.

A base de dados resultante passou por um processo de *feature engineering*, descrito em subseções posteriores, que resultou na obtenção de 93.085 entradas segmentadas

em 3.084 curvas de reversão. Os registros foram posteriormente consolidados em uma base única, padronizada quanto às unidades de medida e à sequência temporal, servindo como entrada para as etapas subsequentes de análise, extração de variáveis e modelagem preditiva.

4.3 Feature Engineering

De acordo com OZDEMIR e SUSARLA (2018), "*feature engineering is the process of transforming data into features that better represent the underlying problem, resulting in improved machine learning performance.*" Em tradução livre, a engenharia de atributos consistiu no processo de transformar dados em variáveis (atributos) que representem de forma mais adequada o fenômeno estudado, permitindo melhor desempenho dos modelos de aprendizado de máquina.

Neste trabalho, a etapa de *feature engineering* compreendeu a organização, segmentação e derivação de variáveis a partir dos registros brutos provenientes dos dispositivos AgNav–Guia. Cada registro de voo foi armazenado em uma linha sequencial, contendo o horário GPS e os parâmetros operacionais descritos na Seção 4.2.

As amostras foram registradas a uma taxa média de 5 Hz quando o sistema de pulverização estava ativado e de 1 a 5 Hz, customizável pelo operador, quando se encontrava desligado. Para garantir uniformidade temporal entre as observações, todas as curvas cujos dados foram gravados em frequências superiores a 1 Hz foram reamostradas para esta taxa, assegurando intervalos temporais constantes entre os registros.

A variável-alvo do estudo correspondeu à variação da altitude GPS (Δalt) em relação ao início de cada curva de reversão individualmente.

A escolha da variável-alvo deveu-se ao fato de que, na operação aeroagrícola, a pulverização é executada a uma altura relativamente constante — geralmente inferior a 20 ft acima do solo — variando conforme o tipo de aplicação realizada. A altitude GPS durante as curvas de reversão varia de acordo com os parâmetros de voo, mas é fortemente influenciada pela altitude em que a manobra se inicia. Dessa forma, a utilização da diferença em relação à altitude inicial padronizou a variável-alvo, permitindo comparações adequadas entre curvas realizadas em diferentes condições operacionais.

A altitude GPS predita pode ser obtida somando-se à variável predita (Δalt) com a altitude de início da curva, conhecida em uma situação real em função da altitude do terreno e da altura de pulverização.

Para identificar as curvas de reversão, adotaram-se critérios de segmentação baseados no comportamento do sistema de pulverização e na variação angular da proa verdadeira. Assim, consideraram-se como trechos de reversão os intervalos entre duas

sequências consecutivas de ativação do sistema de pulverização em que se verificou uma variação de proa igual ou superior a 170° e inferior a 210°.

Essa segmentação permitiu isolar as manobras de interesse, possibilitando a análise dos padrões de variação dos parâmetros durante o reposicionamento e a posterior utilização desses dados na construção de variáveis derivadas e modelos de regressão.

Variáveis derivadas

Após a segmentação das curvas, foram desenvolvidas variáveis adicionais de natureza temporal, operacional, cinemática e geométrica, com o objetivo de representar o comportamento da aeronave ao longo do voo e durante cada manobra.

As variáveis derivadas foram obtidas a partir de cálculos trigonométricos e de medidas de distância haversiniana (entre dois pontos na superfície da Terra). A seguir são apresentadas as categorias de variáveis e a justificativa de sua inclusão na modelagem.

a) Variáveis temporais e operacionais

Tempo de voo total (arquivo): identificou a sequência e a duração completa do voo contido em cada arquivo, geralmente desde a energização da aeronave. O tempo total foi considerado relevante porque, à medida que o voo progride, ocorre uma redução gradual da massa total da aeronave, em decorrência do consumo de combustível.

Essa diminuição de peso afeta diretamente o desempenho, resultando em maior razão de subida e menor velocidade necessária para sustentar o voo. Assim, a altitude alcançada nas curvas pode ser influenciada pela variação de massa. Considerando-se que a massa da aeronave não era registrada pelo equipamento de navegação, esta variável buscou representar essa variação de forma indireta.

Número do voo e tempo de voo por segmento (desde a última decolagem): representam o tempo decorrido desde a última decolagem, marcando os ciclos operacionais típicos das operações aeroagrícolas. Cada segmento costuma a se iniciar após o reabastecimento de insumos de pulverização, o que aumenta o peso da aeronave.

À medida que o voo avança, a aplicação dos insumos e o consumo de combustível reduzem o peso total, alterando a relação peso/potência. A introdução dessa variável teve como objetivo identificar cada abastecimento de insumos, tentando capturar, de forma indireta, a variação de peso decorrente.

Hora do dia: atua como variável relacionada às condições atmosféricas, especialmente temperatura e densidade do ar, fatores que influenciam a sustentação e a eficiência do motor. Como a temperatura ambiente não era registrada diretamente, a hora do dia foi utilizada como indicador indireto das variações atmosféricas médias.

Tempo em curva: corresponde à duração total da manobra de reversão, obtida pela diferença entre o tempo inicial e o final de cada curva. Essa variável permite relacionar

a dinâmica temporal da manobra com a variação de altitude.

Por exemplo, em curvas mais longas, pressupõe-se a ocorrência de mudanças mais suaves de atitude, enquanto curvas rápidas apresentaram maior gradiente vertical.

b) Variáveis cinemáticas

Velocidade inicial da curva: indica o nível de energia cinética da aeronave no início da manobra. A velocidade no ponto de início pode influenciar o raio e a sustentação da curva, trazendo efeitos ao perfil vertical da manobra.

Diferença instantânea para a velocidade inicial da curva: reflete as variações de energia cinética durante a curva, podendo indicar ganhos ou perdas de altitude associados à aceleração ou desaceleração da aeronave.

Aceleração longitudinal: derivada das variações de velocidade sobre o solo, buscou capturar indiretamente as transformações entre energia cinética e potencial, ou seja, entre velocidade e altitude.

c) Variáveis geométricas

Razão de curva (°/s): corresponde à taxa de variação instantânea da proa, podendo influenciar a eficiência da transformação de energia cinética em potencial durante a curva.

Variação angular em relação ao início da curva ($\Delta\psi_{in}$): representa a diferença entre a proa instantânea e a do início da curva. Essa variável descreve a progressão angular da manobra, permitindo distinguir os trechos iniciais (entrada) e finais (saída) da reversão, além de identificar o lado para o qual a curva foi realizada.

Distância em relação ao início da curva: foi calculada pela distância haversiniana entre cada registro e a posição inicial da curva. Possibilita relacionar a trajetória horizontal percorrida com a variação de altitude e reconstruir o perfil tridimensional da manobra.

Ângulo relativo: indica o desvio direcional entre a posição da aeronave e o vetor formado pela posição e a proa de início da curva. Essa variável buscou representar a direção relativa da manobra em relação ao seu início.

Altitude de início da curva: estabeleceu o ponto de referência para as variações verticais, sendo associada às condições de altitude-pressão.

A combinação desses parâmetros derivados permitiu representar de forma mais realista as condições que afetaram a variação de altitude durante as curvas de reversão, atendendo ao princípio de OZDEMIR e SUSARLA (2018) de que o *feature engineering* deve traduzir o conhecimento do domínio físico em representações numéricas adequadas para aprendizado de máquina.

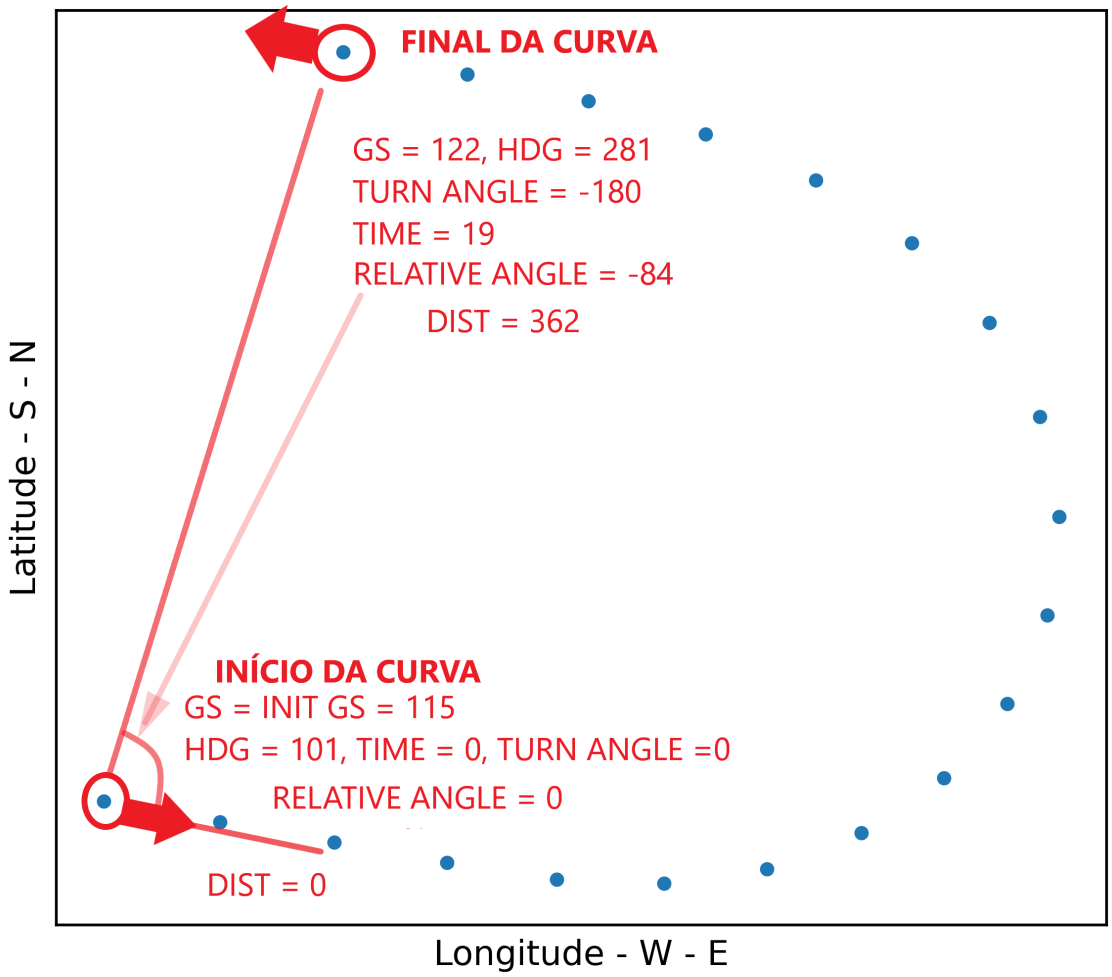

Figura 2 – Exemplo de variáveis em uma curva de reversão (visão do plano horizontal).

A Tabela 3 apresenta as variáveis utilizadas no processo de *feature engineering*, bem como suas respectivas descrições e unidades de medida. Esse detalhamento tem como objetivo garantir transparência metodológica e reprodutibilidade dos experimentos, permitindo a adequada interpretação das características empregadas no treinamento e na avaliação dos modelos preditivos.

Tabela 3 – Descrição dos parâmetros utilizados no modelo preditivo

Parâmetro	**Descrição**
`gs_mph`	Velocidade sobre o solo (mph)
`accel_ms2`	Aceleração longitudinal (m/s^2)
`hour`	Hora do dia (h)
`turn_angle_running`	Ângulo acumulado da curva (°)
`turn_rate_deg_s`	Razão de curva (°/s)
`time_in_turn`	Tempo em curva (s)
`turn_init_spd`	Velocidade inicial da curva (mph)
`dif_init_spd`	Diferença em relação à velocidade inicial (mph)
`dist_from_start_m`	Distância ao início da curva (m)
`flight_number`	Número do voo
`flight_time`	Tempo de voo desde a última decolagem (s)
`total_flight_time`	Tempo total de voo (s)
`rel_angle_deg`	Ângulo relativo (°)
`turn_init_alt`	Altitude inicial da curva (ft)
`dif_init_alt`	Diferença de altitude em relação ao início da curva – variável-alvo (ft)

4.4 Conjuntos de treinamento e teste

Após a consolidação do *dataset* original, contendo as variáveis numéricas resultantes do processo de engenharia de atributos, procedeu-se à divisão dos dados em dois subconjuntos: treinamento e teste. Essa etapa foi realizada utilizando o método *GroupShuffleSplit*, do pacote *scikit-learn*, o qual permitiu a separação dos dados de forma estratificada com base em grupos pré-definidos. No presente estudo, o identificador de grupo empregado foi o campo `turn_id_global`, que representou cada curva de reversão individual.

Essa estratégia impediu que amostras de uma mesma curva fossem divididas entre os conjuntos, evitando vazamento de informação (*data leakage*) e garantindo independência estatística entre as observações de treinamento e teste.

A proporção de divisão adotada foi de aproximadamente 80% dos registros para o conjunto de treinamento e 20% para o conjunto de teste, resultando em 73.622 e 18.423 entradas, respectivamente. Em termos de agrupamentos, o conjunto de treinamento contemplou 2.467 curvas de reversão, enquanto o conjunto de teste incluiu 617 curvas. Essa configuração assegurou uma representação equilibrada da variável-alvo (`dif_init_alt`) em ambos os subconjuntos, preservando a distribuição geral do fenômeno modelado.

A Figura 3 apresenta os histogramas comparativos da variável-alvo — diferença de altitude em relação ao início da curva — para os conjuntos completo, de treinamento e de

teste (da esquerda para a direita), evidenciando que a distribuição se manteve consistente após a separação dos dados, confirmando a adequação do particionamento realizado.

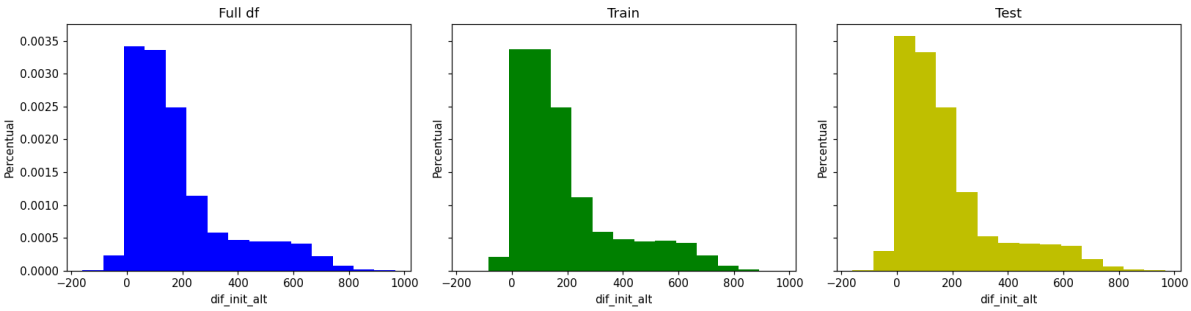

Figura 3 – Distribuição da variável-alvo (`dif_init_alt`) nos conjuntos completo, de treinamento e de teste.

As estatísticas descritivas detalhadas das variáveis consideradas, tanto para o conjunto de treinamento quanto para o conjunto de teste, são apresentadas no Apêndice A. Esses resultados permitem verificar, de maneira quantitativa, a similaridade entre os subconjuntos após o particionamento, reforçando que a divisão adotada preservou as principais características estatísticas dos dados.

4.5 Correlação entre variáveis

A análise de correlação constitui uma etapa fundamental da análise exploratória de dados, pois permite quantificar o grau de associação linear entre pares de variáveis quantitativas e verificar a coerência estatística e física do conjunto de dados.

Dentre as medidas disponíveis, o coeficiente de correlação de Pearson é amplamente empregado por sua interpretação direta e fundamentação estatística consolidada. Esse coeficiente é definido como a razão entre a covariância de duas variáveis e o produto de seus desvios-padrão, assumindo valores no intervalo $[-1, 1]$, em que valores próximos a 1 indicam forte associação linear positiva, valores próximos a -1 indicam forte associação linear negativa e valores próximos de zero indicam ausência de relação linear significativa (Montgomery; Runger, 2018).

Conforme discutido por Montgomery e Runger (2018, p. 143), a correlação de Pearson mede exclusivamente a intensidade da relação linear entre variáveis e não deve ser interpretada como evidência de causalidade. Além disso, relações não lineares, efeitos condicionais ou interações entre múltiplas variáveis podem não ser adequadamente capturados por essa métrica. Ainda assim, a matriz de correlação desempenha papel relevante na validação da consistência dos dados, na identificação de dependências estruturais e no

suporte à interpretação de modelos estatísticos e de aprendizado de máquina, especialmente em sistemas físicos complexos.

No contexto deste trabalho, foi calculada a matriz de correlação de Pearson considerando as variáveis explicativas e a variável-alvo no conjunto de teste, conforme apresentado na Figura 4. A utilização do conjunto de teste nessa análise visa garantir que os padrões estatísticos observados sejam representativos dos dados empregados na avaliação final do desempenho dos modelos, reduzindo o risco de viés associado ao ajuste realizado no conjunto de treinamento.

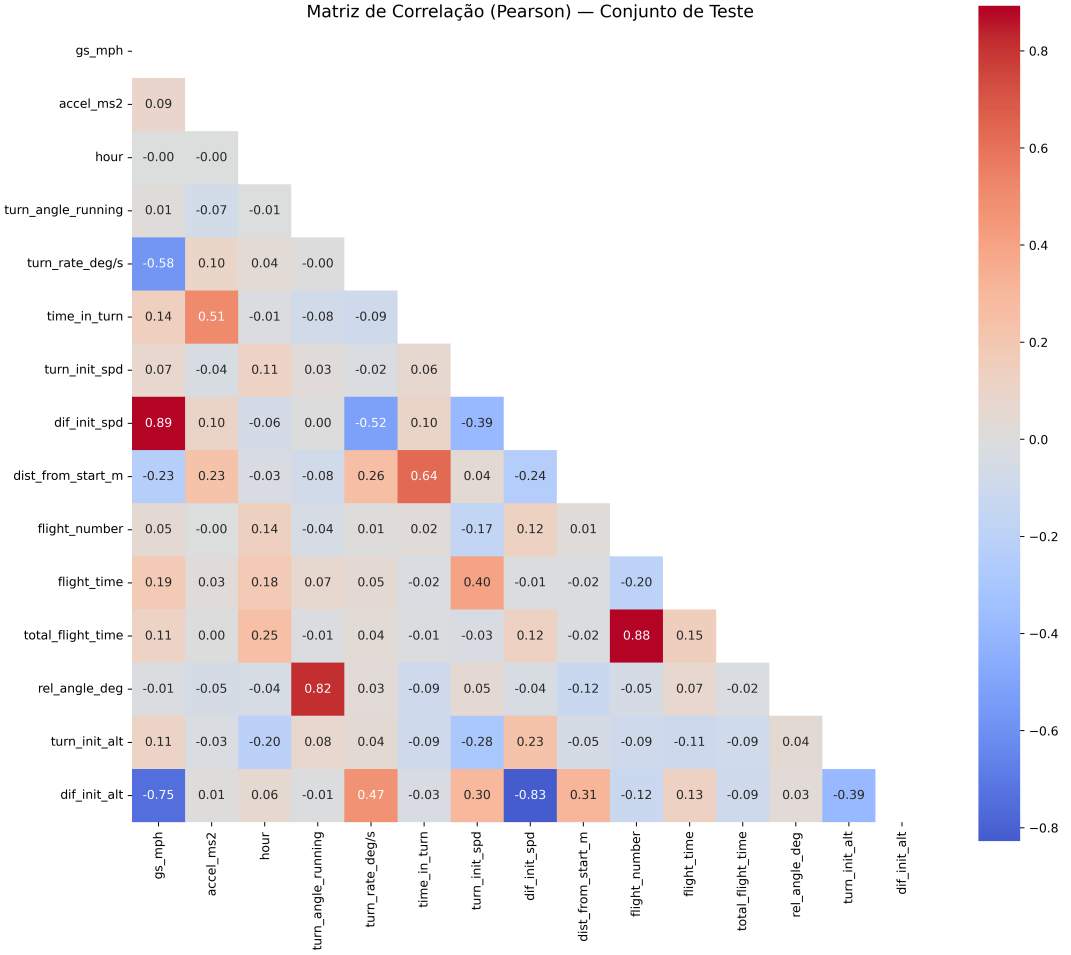

Figura 4 – Matriz de correlação de Pearson entre as variáveis do conjunto de teste.
Fonte: Autor.

A análise da matriz de correlação revela relações estatisticamente consistentes com a dinâmica esperada da manobra de reversão durante operações aeroagrícolas. Observa-se uma correlação negativa elevada entre a velocidade sobre o solo (`gs_mph`) e a diferença de altitude em relação ao início da curva (`dif_init_alt`), variável-alvo, com coeficiente $r = -0{,}75$, indicando que menores velocidades tendem a estar associadas a maiores variações positivas de altitude ao longo da manobra. Esse comportamento é compatível com a conversão de energia potencial em energia cinética durante as manobras.

De forma complementar, a razão de curva (`turn_rate_deg/s`) apresenta correlação positiva moderada com a variável-alvo ($r = 0{,}47$), sugerindo que curvas mais fechadas — caracterizadas por maiores taxas angulares — estejam associadas a maiores variações de altitude. Adicionalmente, a variação da velocidade inicial (`dif_init_spd`) apresenta forte correlação negativa com a diferença de altitude ($r = -0{,}83$), evidenciando, uma vez mais, a relação direta entre alterações na energia cinética e na energia potencial da aeronave ao longo da curva. A própria velocidade inicial da manobra (`turn_init_spd`) também apresenta correlação positiva com a variável-alvo ($r = 0{,}30$), indicando sua influência no regime energético no início da reversão.

Variáveis associadas à geometria e à progressão espacial da manobra, como a distância em relação ao início da curva (`dist_from_start_m`), apresentam correlação positiva moderada com a variável-alvo ($r = 0{,}31$), sugerindo que a variação de altitude tende a se acumular ao longo do desenvolvimento da manobra. Por sua vez, a altitude inicial da curva (`turn_init_alt`) apresenta correlação negativa moderada com a diferença de altitude ($r = -0{,}39$), indicando que curvas iniciadas em níveis mais elevados tendem a apresentar menores variações relativas de altitude ao longo da reversão.

Variáveis associadas à geometria da manobra, como o ângulo acumulado da curva (`turn_angle_running`) e o ângulo relativo (`rel_angle_deg`), exibem forte correlação entre si ($r = 0{,}82$), o que confirma sua consistência geométrica. No entanto, essas variáveis apresentam correlação linear direta pouco significativa com a variável-alvo ($|r| < 0{,}05$), indicando que sua influência sobre a variação de altitude ocorre de forma indireta, mediada por interações com velocidade, taxa de curva e duração da manobra.

Por sua vez, variáveis temporais e operacionais, como `flight_number` ($r = -0{,}12$), `flight_time` ($r = 0{,}13$) e `total_flight_time` ($r = -0{,}09$), apresentam correlações de baixa magnitude com a variável-alvo, sugerindo que seu papel no modelo está associado a efeitos contextuais e a interações de ordem superior, e não a relações lineares diretas.

4.6 Aspectos computacionais

A evolução das técnicas computacionais permitiu avanços significativos na análise de dados, especialmente em áreas que demandam alta eficiência e precisão na construção de modelos preditivos. Ferramentas de aprendizado de máquina desempenharam papéis relevantes nesse contexto, possibilitando a automação de tarefas complexas e a otimização do desempenho em diferentes domínios.

Com o uso de bibliotecas especializadas, tornou-se possível implementar e avaliar algoritmos de forma eficiente, reduzindo o tempo de processamento e facilitando a integração de soluções em *pipelines* de dados robustos. Nesse contexto, destacou-se a biblioteca PyCaret PYCARET (2023), que combinou simplicidade de uso com uma abordagem abrangente

para análise preditiva.

4.6.1 PyCaret

PyCaret é uma biblioteca *open-source* desenvolvida em Python com o propósito de simplificar e automatizar fluxos de trabalho de aprendizado de máquina, permitindo a construção de modelos com poucas linhas de código. A ferramenta possibilitou a aplicação de algoritmos que abrangem desde técnicas clássicas, como regressão linear e árvores de decisão, até métodos modernos de *ensemble* e *boosting*, integrados de forma a facilitar a experimentação e a otimização de desempenho PYCARET (2023).

No presente trabalho, a utilização da PyCaret seguiu o fluxo metodológico sumarizado na Tabela 4, que descreve os principais parâmetros empregados na configuração do ambiente experimental.

Tabela 4 – Principais parâmetros utilizados na configuração da PyCaret

Parâmetro	**Valor**	**Definição**
`data`	`df_sub`	Conjunto de dados utilizado, contendo 30% dos grupos amostrados por `turn_id_global`.
`target`	(Δalt) (`dif_init_alt`)	Variável-alvo a ser prevista pelo modelo, definida como a diferença de altitude em relação ao início da curva.
`fold_strategy`	`groupkfold`	Estratégia de validação cruzada que mantém os registros do mesmo grupo (`turn_id_global`) na mesma partição.
`fold_groups`	`turn_id_global`	Identificador de grupo utilizado para segmentação e controle de independência entre conjuntos de treino e validação.
`ignore_features`	`['turn_id_global']`	Colunas excluídas da modelagem, não utilizadas como preditoras.
`session_id`	6	Identificador de sessão utilizado para garantir reprodutibilidade dos experimentos.
`remove_multicollinearity`	`True`	Ativou a remoção automática de variáveis altamente correlacionadas, evitando redundância entre preditores.
`multicollinearity_threshold`	0.9	Definiu o limite de correlação acima do qual variáveis redundantes foram eliminadas, mantendo apenas um dos atributos correlatos.
`sort`	`'MAE'`	Métrica utilizada para ranquear os modelos com base no Erro Absoluto Médio (*Mean Absolute Error*).

Primeiramente, foi realizada uma amostragem de 30% dos grupos utilizando o método *GroupShuffleSplit*, garantindo independência entre os conjuntos de treinamento e validação no nível dos grupos definidos por `turn_id_global`. A opção por empregar apenas uma fração dos dados de treinamento para aplicação na PyCaret, teve como objetivo reduzir o tempo computacional, permitindo que apenas os modelos mais promissores fossem posteriormente avaliados no conjunto completo.

O uso do *GroupShuffleSplit* assegurou que segmentos completos de curva fossem mantidos, preservando a natureza sequencial dos registros. Dessa forma, evitou-se que o treinamento ocorresse sobre condições temporais que pudessem introduzir dependência ou sobreposição com o conjunto de teste.

O *pipeline* considerou variáveis definidas conceitualmente como exógenas antes da etapa de seleção de *features* e modelagem. Essas variáveis foram integradas ao conjunto de preditores e submetidas aos mesmos procedimentos automáticos de pré-processamento da PyCaret, incluindo normalização, remoção de multicolinearidade e seleção de atributos (parâmetros `remove_multicollinearity=True` e `multicollinearity_threshold=0.9`).

Essa abordagem mostrou-se associada a ganhos de desempenho preditivo, refletidos em valores mais elevados do coeficiente de determinação (R^2) e na redução do erro médio absoluto (MAE) nos modelos selecionados.

Por fim, foi realizada a comparação automática dos modelos utilizando o comando `compare_models(sort="MAE")`, que selecionou os algoritmos com melhor desempenho na métrica de erro absoluto médio para posterior refinamento e avaliação aprofundada.

4.7 Métricas de Avaliação

A avaliação do desempenho de um modelo de aprendizado de máquina depende diretamente do tipo de problema tratado. Nos casos de regressão, o objetivo consiste em estimar valores contínuos de uma variável de interesse a partir de um conjunto de preditores. Assim, a análise da qualidade do modelo envolve métricas específicas que mensuram o grau de proximidade entre os valores preditos ($\hat{Y}$) e os valores observados (Y).

As métricas mais utilizadas nesse contexto são o Erro Absoluto Médio (MAE), a Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2).

Essas métricas, amplamente conhecidas academicamente, são definidas a seguir de acordo com a documentação oficial da biblioteca de aprendizado de máquina Scikit-learn (Scikit-learn Developers, 2025).

4.7.1 Erro Absoluto Médio (MAE)

O Erro Absoluto Médio (MAE) representa a média das diferenças absolutas entre os valores reais e os valores previstos pelo modelo. Sua formulação matemática é dada por:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{4.1}$$

Em termos interpretativos, o MAE expressa o erro médio absoluto cometido nas previsões, indicando, em média, o quanto as estimativas do modelo se distanciam dos valores reais observados. Por ser calculado com valores absolutos, essa métrica atribui o mesmo peso a todos os erros, não penalizando discrepâncias grandes com maior intensidade.

Outra característica importante é que o MAE é expresso na mesma unidade da variável predita — no contexto deste trabalho, uma altura medida em *feet* (ft) —, o que facilita sua interpretação prática. Em síntese, valores menores de MAE correspondem a melhor desempenho preditivo.

4.7.2 Raiz do Erro Quadrático Médio (RMSE)

O Erro Quadrático Médio (MSE) e sua raiz quadrada, o RMSE (*Root Mean Squared Error*), são métricas que avaliam o erro médio, mas com uma ponderação diferenciada, pois elevam os resíduos ao quadrado antes da média. Sua expressão é:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{4.2}$$

A principal diferença entre o RMSE e o MAE é que o RMSE penaliza mais fortemente erros de maior magnitude, tornando-se mais sensível à presença de *outliers*. Essa característica é desejável quando se busca um modelo que minimize grandes desvios individuais. Assim como o MAE, o RMSE é medido na mesma unidade da variável dependente, o que permite interpretações diretas.

Em geral, valores menores de RMSE indicam melhor ajuste entre as previsões do modelo e os dados observados, com uma ênfase maior na redução de grandes erros.

4.7.3 Coeficiente de Determinação (R^2)

O Coeficiente de Determinação (R^2) é uma métrica adimensional que expressa a proporção da variabilidade dos dados explicada pelo modelo de regressão. É definido pela equação:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{4.3}$$

em que $\bar{Y}$ representa a média dos valores observados.

O valor de R^2 varia entre 0 e 1, podendo eventualmente assumir valores negativos quando o modelo apresenta desempenho inferior ao de uma simples média. Um R^2 igual a 1 indica um ajuste perfeito, no qual o modelo explica integralmente a variância da variável dependente. Por outro lado, um R^2 próximo de 0 evidencia que o modelo praticamente não explica a variabilidade dos dados.

O R^2 é interpretado como uma medida da capacidade explicativa do modelo, sendo amplamente utilizado em conjunto com as métricas de erro (MAE e RMSE) para uma avaliação mais abrangente.

5 RESULTADOS

Este capítulo apresenta e analisa os resultados obtidos a partir da aplicação dos modelos desenvolvidos, permitindo ao leitor compreender o desempenho das abordagens avaliadas no contexto do estudo. São discutidas as métricas utilizadas para a avaliação dos modelos, bem como a distribuição dos erros e os padrões observados nos dados de teste. O capítulo explora a relação entre o desempenho preditivo e as características operacionais representadas pelas variáveis de entrada, destacando situações nas quais o modelo apresenta maior ou menor acurácia. A análise dos resultados fornece subsídios para a compreensão das limitações do modelo e de seu potencial de aplicação no apoio a análises voltadas à segurança de voo.

5.1 Seleção dos modelos mais promissores

Para a seleção dos modelos mais promissores, utilizou-se uma fração do conjunto de treinamento, conforme descrito na Seção 4.6.1. Posteriormente, esse modelos foram avaliados pelas métricas: Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2), aplicados ao conjunto de validação, além do tempo de treinamento.

O ambiente experimental foi estruturado, em uma CPU, *Central Porcessing Unit*, com o *framework* PyCaret, que automatiza as etapas de comparação entre algoritmos, cálculo das métricas de desempenho e validação cruzada. Durante a execução, o PyCaret aplicou um conjunto de algoritmos lineares e não lineares, exibindo em uma tabela comparativa o desempenho médio de cada modelo para um conjunto denominado de validação, correspondente a 25% dos dados fornecidos à biblioteca.

A partir dessa lista inicial, foram selecionados para análise detalhada os modelos com melhor desempenho dentro de cada grupo — lineares e não lineares — com o intuito de discutir tanto a precisão quanto a interpretabilidade das abordagens.

Os resultados obtidos estão apresentados na Tabela 5. O modelo *Extra Trees Regressor* apresentou o melhor desempenho geral, com MAE = 29,52, RMSE = 44,47 e R^2 = 0,94. O *CatBoost Regressor* exibiu métricas muito próximas (MAE = 30,91 e R^2 = 0,94), embora com tempo de treinamento consideravelmente superior (TT = 9136 s), o que impacta sua aplicabilidade prática em cenários de análise operacional.

Em seguida, o *Random Forest Regressor* e o *Light Gradient Boosting Machine (LightGBM)* apresentaram resultados ligeiramente inferiores ($R^2 \approx$ 0,93), mas ainda consistentes, confirmando a adequabilidade dos métodos baseados em árvores de decisão

para dados complexos e de alta variabilidade, como os registrados nas curvas de reversão.

Por outro lado, os modelos lineares — *Linear Regression* e *Ridge Regression* — obtiveram desempenho substancialmente inferior, com MAE e RMSE próximos de 65 e 85, respectivamente, e $R^2 = 0{,}78$. Esses resultados indicam que tais modelos possuem capacidade limitada de representar relações não lineares e interações múltiplas entre variáveis, como as esperadas nas manobras avaliadas. Embora os modelos lineares apresentem maior interpretabilidade e reduzido custo computacional, sua explicabilidade é restrita diante da complexidade estrutural dos dados.

A avaliação revelou que os modelos não lineares, baseados em árvores de decisão, mostraram-se mais adequados ao problema proposto, equilibrando precisão, estabilidade e capacidade de generalização.

Tabela 5 – Resultados dos modelos de regressão no PyCaret

Modelo	**MAE**	**RMSE**	R^2	**TT (s)**
Extra Trees Regressor	29,5231	44,4671	0,9404	878
CatBoost Regressor	30,9129	44,6894	0,9399	9136
Random Forest Regressor	31,8668	49,2443	0,9261	2233
Light Gradient Boosting Machine	32,7949	47,8500	0,9311	883
Ridge Regression	64,7836	84,7965	0,7840	34
Linear Regression	64,7837	84,7963	0,7840	46

Fonte: Autor.

5.2 Avaliação final dos modelos não lineares no conjunto de teste

Após a etapa de seleção e validação dos modelos mais promissores no conjunto de treinamento, procedeu-se à avaliação final dos modelos não lineares no conjunto de teste. Essa etapa tem como objetivo comparar o desempenho dos modelos em dados não utilizados durante o treinamento, fornecendo uma estimativa robusta de sua capacidade de generalização.

Foram avaliados os seguintes algoritmos: *CatBoost Regressor*, *Light Gradient Boosting Machine (LightGBM)*, *Extra Trees Regressor* e *Random Forest Regressor*. Todos os modelos foram aplicados ao mesmo conjunto de teste, composto por 18.423 entradas em 617 curvas de reversão, garantindo condições equivalentes de comparação.

Foi considerada uma métrica complementar baseada no percentil 95 do erro absoluto médio por curva, doravante denominada $P95$ (MAE). Diferentemente das métricas globais, como o MAE e o RMSE, que são calculadas diretamente sobre todas as entradas individuais do conjunto de teste, o $P95$ (MAE) é obtido a partir da agregação das observações por curva de reversão. Para tal, calcula-se inicialmente o erro absoluto médio (MAE) para cada curva de forma independente, considerando todas as amostras temporais associadas

a essa manobra. Em seguida, determina-se o percentil 95 da distribuição desses valores de MAE por curva. Assim, o $P95$ (MAE) representa o valor abaixo do qual se encontram 95% dos erros médios observados entre as curvas avaliadas, fornecendo uma estimativa do limite superior do erro típico no regime operacional mais frequente.

A métrica $P95$ (MAE) permite avaliar a consistência do desempenho do modelo ao longo das diferentes curvas de reversão, reduzindo a influência de curvas atípicas com elevado número de amostras ou de erros pontuais localizados. Dessa forma, o $P95$ (MAE) complementa a análise baseada no MAE global, oferecendo uma visão mais alinhada ao comportamento do modelo no nível da manobra, que constitui a unidade operacional de interesse neste estudo.

O custo computacional foi analisado de forma complementar, considerando o tempo total de treinamento, definido como a soma do tempo de validação cruzada e do tempo de treinamento final do modelo ajustado com 100% do conjunto de treinamento. Essa métrica permite avaliar a viabilidade prática dos modelos no ambiente experimental adotado.

Os resultados consolidados da avaliação no conjunto de teste encontram-se apresentados na Tabela 6.

Tabela 6 – Desempenho final dos modelos não lineares no conjunto de teste

Modelo	**MAE**	**RMSE**	R^2	$P95$ (MAE)	**TT (s)**
CatBoost Regressor	21,4212	34,3516	0,9623	44,014	91,28
LightGBM	21,8031	35,2534	0,9603	45,216	52,04
Extra Trees Regressor	22,0153	37,2526	0,9557	46,162	71,50
Random Forest Regressor	23,9152	40,3240	0,9481	50,662	212,48

Fonte: Autor.

A consistência observada entre o MAE e o $P95$ (MAE) reforça a robustez dos modelos com melhor desempenho, indicando que a redução do erro médio não ocorre às custas de degradação nos casos mais desfavoráveis. Esse comportamento sugere melhor capacidade de generalização.

5.3 Seleção do modelo final

A seleção do modelo final baseou-se na análise conjunta do desempenho preditivo no conjunto de teste, considerando tanto uma métrica de erro central quanto uma medida associada ao comportamento típico da maior parte das curvas avaliadas. Em particular, foram adotados como critérios principais o Erro Médio Absoluto (MAE) e o percentil 95 do erro absoluto médio por curva ($P95$ (MAE)).

Conforme apresentado na Tabela 6, o *CatBoost Regressor* apresentou os melhores resultados globais em todas as métricas, destacando-se o menor MAE de 21,42 e o menor

valor de $P95$ (MAE) de 44,01. Esses resultados indicam não apenas maior precisão média, mas também menor erro no intervalo que abrange 95% das curvas avaliadas. O *LightGBM* apresentou desempenho muito próximo, com MAE de 21,80 e $P95$ (MAE) de 45,22, seguido pelo *Extra Trees Regressor*. O *Random Forest Regressor* apresentou desempenho inferior em ambas as métricas consideradas.

A coerência observada entre a ordenação dos modelos segundo o MAE e o $P95$ (MAE) reforça a consistência dos resultados, sugerindo que os modelos com menor erro médio também tendem a apresentar menor erro no regime mais comum de operação. Esse comportamento é particularmente relevante para aplicações operacionais, nas quais se busca desempenho estável e previsível na grande maioria dos casos.

Em relação ao custo computacional, embora diferenças no tempo total de treinamento tenham sido observadas entre os modelos, essas não se mostraram relevantes no contexto experimental adotado. Os experimentos foram conduzidos utilizando uma GPU L4 disponibilizada pela plataforma Google Colab, o que contribuiu para reduzir o tempo de processamento e tornar as diferenças de tempo total de treinamento pouco significativas para fins práticos nesta etapa do estudo. Ressalta-se que essa análise é válida para o ambiente computacional considerado.

Cabe destacar ainda que, diferentemente da avaliação inicial realizada com o *framework* PyCaret, o treinamento final dos modelos foi conduzido utilizando a totalidade do conjunto de treinamento. Observou-se uma redução expressiva do MAE no conjunto de teste em comparação com a triagem inicial, sugerindo que o aumento do volume de dados empregados no ajuste contribuiu positivamente para o aprendizado das relações subjacentes ao fenômeno analisado, quando acompanhado de uma estratégia de validação apropriada.

Dessa forma, considerando o desempenho superior em termos de MAE e $P95$ (MAE) no conjunto de teste, bem como a ausência de restrições computacionais relevantes no ambiente experimental adotado, o *CatBoost Regressor* foi selecionado como modelo final para as análises subsequentes.

A viabilidade do modelo *CatBoost Regressor* pode ser avaliada não apenas a partir das métricas globais de erro, mas também pela relação entre esses erros e a escala da variável resposta no conjunto de teste. A variável predita apresenta mediana de 125 e média de aproximadamente 176 ft, com ampla variabilidade entre curvas. Nesse contexto, o MAE do modelo, da ordem de 21 ft, corresponde a cerca de 12% da média e 17% da mediana da variável alvo, enquanto o percentil 95 do MAE por curva, aproximadamente 44 ft, permanece abaixo de 25% da média observada.

A Figura 5 apresenta a distribuição do MAE por curva no conjunto de teste para os quatro modelos avaliados. Muito embora o CatBoost tenha apresentado melhores

resultados, em todos modelos ficou evidenciado que a grande maioria das curvas apresenta erro substancialmente inferior ao percentil 95, o que reforça a consistência do desempenho dos modelos no regime mais frequente de operação. Observa-se ainda a presença de distribuições com cauda longa, associada a um número percentualmente reduzido de curvas com erro elevado.

Complementarmente, a Figura 6 ilustra, para o modelo CatBoost, dois exemplos representativos para cada conjunto de curvas com diferentes níveis de erro: menores, próximas ao comportamento médio, próximas ao percentil 95 e com maior erro. Além da comparação visual entre valores observados e preditos, são apresentadas métricas individuais de MAE, RMSE e R^2 para cada curva, permitindo avaliar como o desempenho global do modelo se manifesta em situações específicas.

Em conjunto, essas análises indicam que o modelo CatBoost apresentou os melhores resultados para este estudo e que seu desempenho foi adequado e consistente para a predição da diferença de altitude durante curvas de reversão, com erros relativamente pequenos quando comparados à escala da variável de interesse.

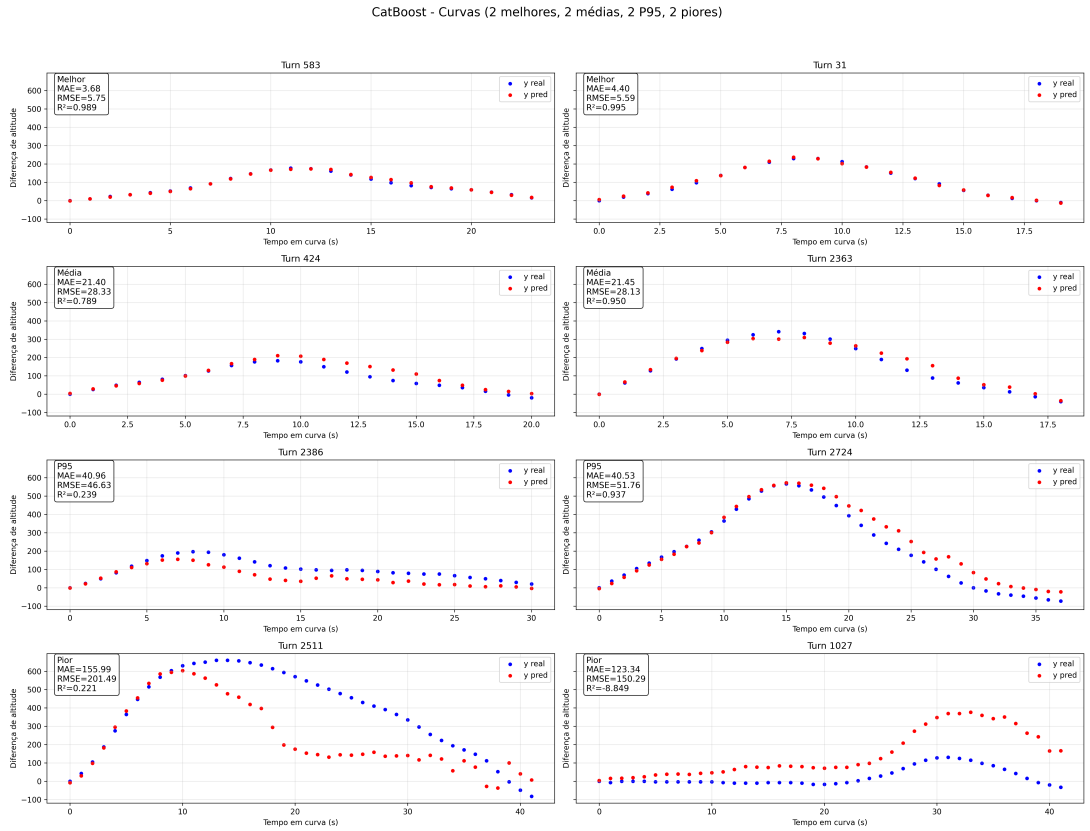

Figura 6 – Exemplos de comparação entre valores observados e preditos para o modelo CatBoost

Fonte: Autor.

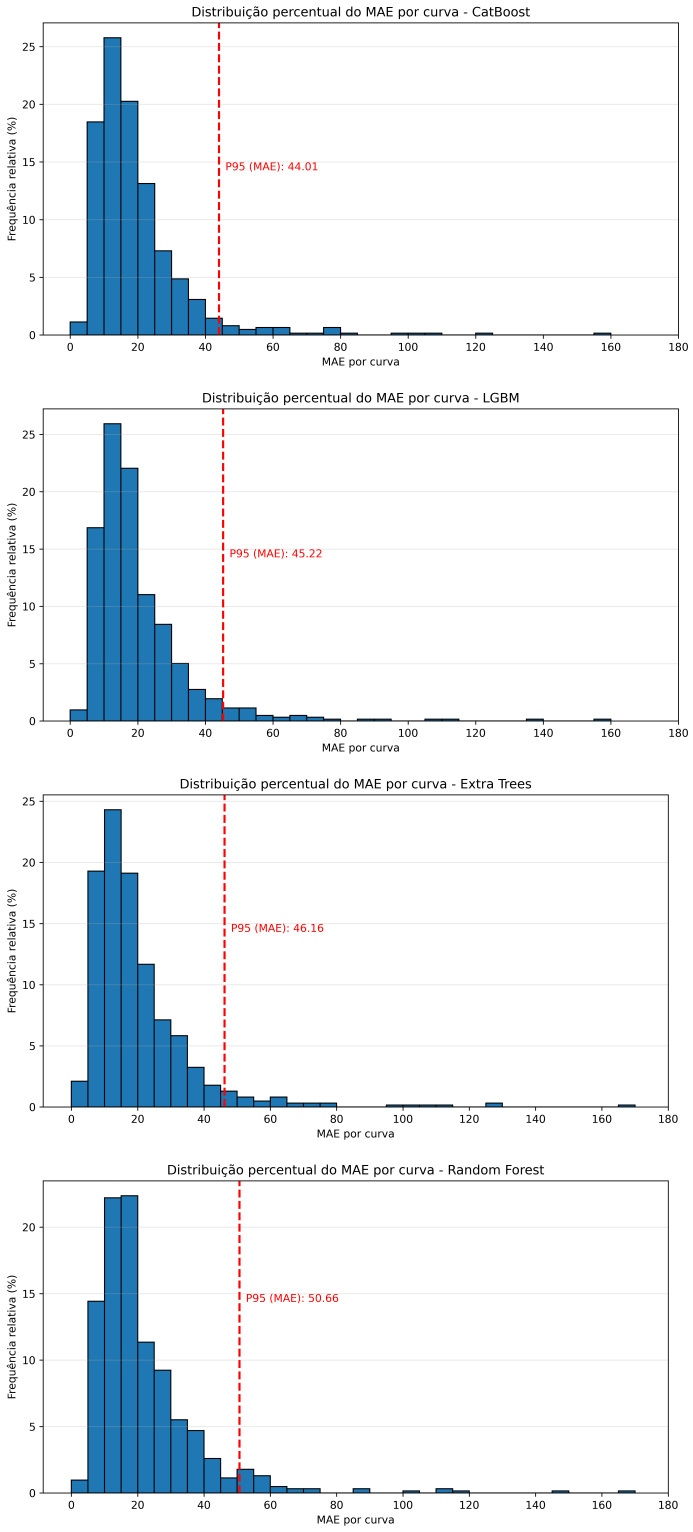

Figura 5 – Distribuição do erro absoluto médio (MAE) por curva no conjunto de teste para os modelos avaliados.

Fonte: Autor.

Apesar de o CatBoost ter apresentado o melhor desempenho neste estudo, sua vantagem não foi substancial em relação aos demais modelos baseados em árvores de decisão, sugerindo que diferentes algoritmos dessa família podem alcançar resultados comparáveis sob condições similares. Assim, é importante considerar que, ao se aplicar a mesma abordagem a outros conjuntos de dados — por exemplo, de aeronaves distintas, com perfis operacionais diferentes — pode ocorrer que outro modelo apresente desempenho superior, reforçando a necessidade de validação empírica caso a caso.

5.4 Análise de explicabilidade do modelo

Modelos não lineares baseados em árvores de decisão costumam apresentar alto desempenho preditivo em problemas complexos, mas sua estrutura interna dificulta a interpretação direta das relações entre variáveis de entrada e a saída estimada. Em aplicações associadas a sistemas físicos, como a dinâmica de voo em manobras aeroagrícolas, essa limitação torna a análise de explicabilidade uma etapa relevante do processo de modelagem.

Neste trabalho, o modelo selecionado (*CatBoost Regressor*) combina múltiplas árvores de decisão e interações não lineares entre variáveis explicativas. Embora isso seja fundamental para capturar a complexidade do fenômeno analisado, dificulta compreender, de forma direta, a contribuição individual de cada variável para a estimativa da diferença de altitude ao longo das curvas de reversão, o que motiva o uso de técnicas específicas de explicabilidade.

A análise de explicabilidade tem como objetivo avaliar se o modelo fundamenta suas previsões em relações fisicamente plausíveis e compatíveis com a dinâmica da manobra, bem como investigar o comportamento do desempenho ao nível de cada curva, e não apenas por métricas globais agregadas. Dessa forma, a explicabilidade complementa a avaliação quantitativa, fornecendo subsídios para a interpretação técnica dos resultados e a validação conceitual do modelo.

5.4.1 Explicabilidade baseada em valores SHAP

A explicabilidade de modelos de aprendizado de máquina tem sido amplamente abordada por métodos fundamentados na teoria dos jogos cooperativos, dentre os quais se destaca o SHAP (*SHapley Additive exPlanations*). O método SHAP fornece uma estrutura unificada para interpretar predições de modelos complexos, atribuindo a cada variável explicativa uma contribuição quantitativa associada à saída do modelo, de forma consistente e aditiva (Lundberg; Lee, 2017).

No contexto de modelos preditivos, os valores SHAP (*SHAP values* ou *SHAP numbers*) quantificam a contribuição marginal de cada variável para a predição de uma

observação recall, considerando todas as combinações possíveis de variáveis. A predição do modelo pode ser expressa como a soma do valor esperado da saída e das contribuições individuais associadas a cada variável, segundo a decomposição:

$$\hat{y} = \mathbb{E}[\hat{y}] + \sum_{j=1}^{p} \phi_j,$$

em que ϕ_j representa o valor SHAP da variável j. Valores positivos de ϕ_j indicam contribuições que elevam a predição em relação ao valor médio do modelo, enquanto valores negativos indicam contribuições que reduzem a predição.

Uma característica central do SHAP é a possibilidade de análise tanto em nível local quanto global. Em nível local, os valores SHAP permitem interpretar predições individuais, fornecendo explicações detalhadas para instâncias específicas. Em nível global, a importância das variáveis pode ser avaliada a partir da média do valor absoluto dos valores SHAP ao longo do conjunto de dados, resultando em uma métrica comparável entre variáveis e alinhada ao impacto efetivo destas sobre as predições do modelo (Lundberg; Lee, 2017).

Para modelos baseados em árvores de decisão, como o *CatBoost Regressor*, o cálculo dos valores SHAP pode ser realizado de forma eficiente por meio de implementações específicas, como o *TreeExplainer*, que exploram a estrutura do modelo para reduzir o custo computacional sem comprometer a fidelidade das explicações (SHAP Developers, 2024). Essa abordagem é particularmente adequada ao presente estudo, no qual o modelo incorpora interações não lineares e dependências complexas entre variáveis associadas à dinâmica de voo.

Nas análises subsequentes, o método SHAP é empregado para investigar a importância global das variáveis explicativas e para interpretar o comportamento do modelo em observações individuais, estabelecendo uma base sólida para a análise de explicabilidade no contexto das curvas de reversão avaliadas.

5.4.2 Importância global das variáveis segundo SHAP

A partir dos valores SHAP calculados para o modelo CatBoost, foi realizada uma análise de importância global das variáveis explicativas, considerando a média do valor absoluto SHAP ao longo de todas as observações do conjunto de teste. Essa métrica, quantifica o impacto médio de cada variável sobre a predição do modelo, permitindo uma comparação direta entre variáveis independentemente do sinal da influência.

A Figura 7 apresenta o gráfico de barras correspondente à importância global das variáveis segundo o critério da média do valor absoluto SHAP, ordenadas de forma decrescente. Variáveis posicionadas no topo do gráfico exercem maior influência média sobre as predições do modelo, enquanto aquelas posicionadas na base apresentam impacto global reduzido.

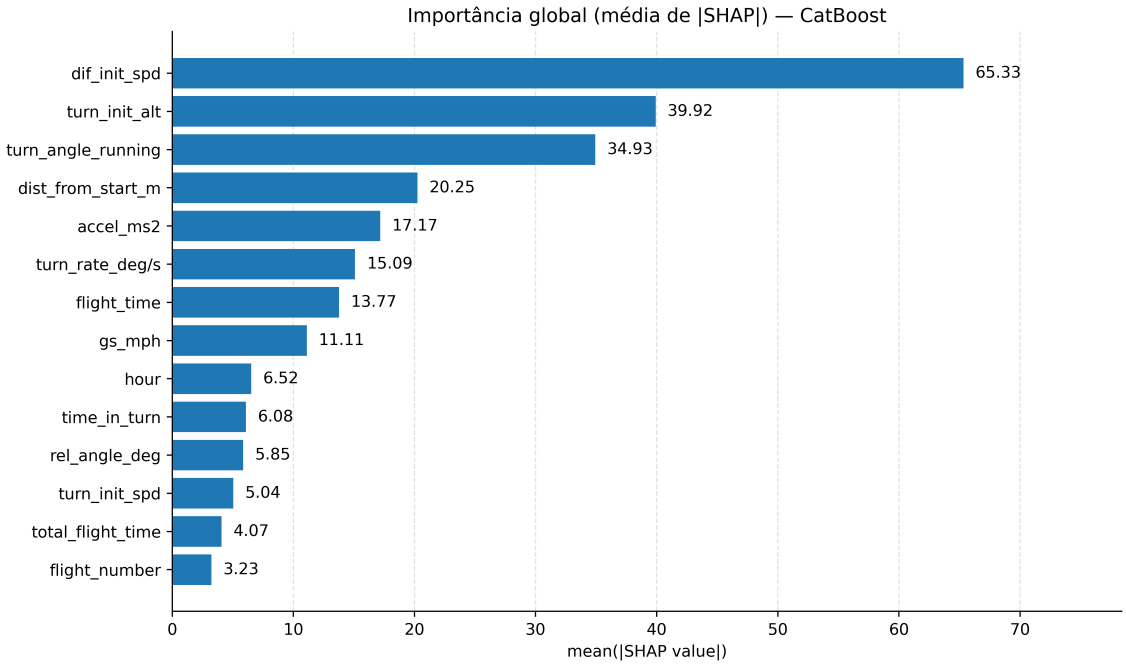

Figura 7 – Importância global das variáveis segundo a média do valor absoluto dos SHAP values para o modelo CatBoost.

Fonte: Autor.

Observa-se que a variável `dif_init_spd` apresenta a maior importância global, com valor médio de $|\phi| = 65{,}33$, indicando que a variação em relação à velocidade inicial exerce impacto significativo sobre as predições do modelo. Em seguida, destacam-se as variáveis `turn_init_alt` ($|\phi| = 39{,}92$), comentada ao final desta subseção, e `turn_angle_running` ($|\phi| = 34{,}93$), ambas associadas às condições iniciais e à geometria da manobra.

Variáveis relacionadas à progressão espacial e dinâmica da curva, como `dist_from_start_m` ($|\phi| = 20{,}25$), `accel_ms2` ($|\phi| = 17{,}17$) e `turn_rate_deg/s` ($|\phi| = 15{,}09$), apresentam importância intermediária, sugerindo contribuição consistente ao longo das curvas, porém com menor peso médio quando comparadas às variáveis dominantes. Esse comportamento indica que tais variáveis atuam de forma complementar, interagindo com as condições iniciais e com a geometria da manobra.

Por outro lado, variáveis temporais e operacionais, como `flight_time` ($|\phi| = 13{,}77$), `hour` ($|\phi| = 6{,}52$), `total_flight_time` ($|\phi| = 4{,}07$) e `flight_number` ($|\phi| = 3{,}23$), apresentam valores médios SHAP reduzidos, indicando influência global limitada sobre as predições. Esse resultado sugere que tais variáveis não atuam como determinantes diretos da variação de altitude, mas podem contribuir por meio de interações de ordem superior ou em contextos específicos da operação.

De modo geral, a análise de importância global baseada em valores SHAP pode ser confrontada com a matriz de correlação, apresentada na Seção 4.5, e com as relações aerodinâmicas conhecidas.

Enquanto a correlação de Pearson avalia exclusivamente associações lineares entre pares de variáveis, os valores SHAP capturam contribuições não lineares e interações complexas entre atributos, permitindo identificar variáveis que exercem influência relevante mesmo na ausência de correlação linear elevada com a variável-alvo.

Nesse contexto, destaca-se a elevada importância atribuída à variável `turn_init_alt`, a qual apresenta correlação linear moderada com `dif_init_alt` ($r \approx -0{,}39$), mas surge como uma das variáveis mais influentes segundo a métrica SHAP. Entretanto, essa variável apresenta uma amplitude de variação da ordem de 500 ft, conforme evidenciado no Apêndice A, considerada muito pequena, do ponto de vista aerodinâmico, para exercer tamanha influência no modelo.

Esse resultado sugere que sua relevância não decorre de um efeito aerodinâmico direto, mas, possivelmente, de seu papel como variável proxy de condições operacionais ou padrões recorrentes de manobra em faixas de pulverização contíguas, ou seja, geograficamente correlacionadas.

5.4.3 Modelo alternativo com reparametrização da altitude inicial

Conforme evidenciado na análise de explicabilidade do modelo original, baseada em valores SHAP, a variável `turn_init_alt` apresentou uma importância global superior ao esperado do ponto de vista físico-aerodinâmico. Tal resultado levantou a hipótese de que o modelo poderia estar, ao menos parcialmente, aprendendo padrões operacionais específicos associados a faixas de pulverização contíguas, nas quais a altitude inicial apresenta variação limitada e está fortemente correlacionada com a localização geográfica da operação, em vez de refletir um efeito causal direto sobre a aerodinâmica da manobra.

Com o objetivo de investigar essa hipótese e reduzir a sensibilidade do modelo a pequenas variações de altitude, potencialmente associadas a efeitos contextuais, foi construído um modelo alternativo, equivalente ao modelo original em termos de arquitetura, hiperparâmetros e conjunto de variáveis, diferindo apenas na reparametrização da variável `turn_init_alt`.

Nesse modelo, a altitude inicial foi transformada para uma escala em milhares de pés, por meio de arredondamento para o milhar mais próximo, reduzindo sua resolução e, consequentemente, sua capacidade de capturar diferenças locais sutis. Essa transformação resulta em uma variável com baixa variabilidade, conforme indicado por suas estatísticas descritivas (média de 1,11, desvio-padrão de 0,32 e valores mínimo e máximo de 1 e 2, respectivamente).

A Tabela 7 apresenta o desempenho dos modelos alternativos no conjunto de teste, permitindo comparação direta com o modelo original.

Tabela 7 – Desempenho dos modelos com `turn_init_alt` reparametrizados (em milhares de ft)

Modelo	**MAE**	**RMSE**	R^2	$P95$ (MAE)
CatBoost Regressor (2)	24,5192	40,3866	0,9479	47,002
Extra Trees Regressor (2)	24,7281	42,4152	0,9425	50,997
LightGBM (2)	24,7865	41,0673	0,9461	49,439
Random Forest Regressor (2)	26,1038	44,9593	0,9354	54,653
(2) - reparametrizado em milhares de ft				

Fonte: Autor.

Novamente, o CatBoost apresentou o melhor desempenho entre os modelos avaliados, mantendo a liderança em todas as métricas em ambos os cenários.

Observa-se que a reparametrização da altitude inicial resultou em uma degradação moderada do desempenho preditivo, relativamente uniforme em todos os modelos, com aumento do MAE, RMSE, P95 (MAE) e redução do coeficiente de determinação quando comparados ao modelo original.

Embora o CatBoost tenha apresentado a maior degradação percentual em MAE, RMSE e R^2 após a reparametrização, ele foi o modelo com a menor variação percentual em P95 (MAE). Esse resultado indica que, entre as curvas mais comuns, o limite superior do erro por curva foi menos sensível à reparametrização, ainda que as métricas globais apontem para uma deterioração mais pronunciada do desempenho médio.

A Tabela 8 apresenta a variação percentual das métricas de desempenho para cada modelo após a reparametrização da altitude.

Tabela 8 – Variação percentual das métricas após reparametrização (referência: modelos originais)

Modelo	**MAE**	**RMSE**	R^2	$P95$ (MAE)
CatBoost Regressor	+14,46%	+17,57%	-1,50%	+6,79%
Extra Trees Regressor	+12,32%	+13,86%	-1,38%	+10,47%
LightGBM	+13,68%	+16,49%	-1,48%	+9,34%
Random Forest Regressor	+9,15%	+11,50%	-1,34%	+7,88%

Fonte: Autor.

Esse comportamento indica que, no conjunto de dados analisado, a informação contida na altitude inicial em sua forma original contribui para a redução do erro médio, ainda que parte dessa contribuição possa estar associada a padrões específicos do contexto operacional.

A Figura 8 apresenta a importância global das variáveis segundo a média do valor absoluto SHAP para o modelo CatBoost. Nota-se uma redução expressiva da importância atribuída à variável `turn_init_alt` ($|\phi| = 3{,}39$), que passa a ocupar uma posição marginal no ranking de contribuições, confirmando que a transformação aplicada limitou sua influência direta sobre as predições do modelo.

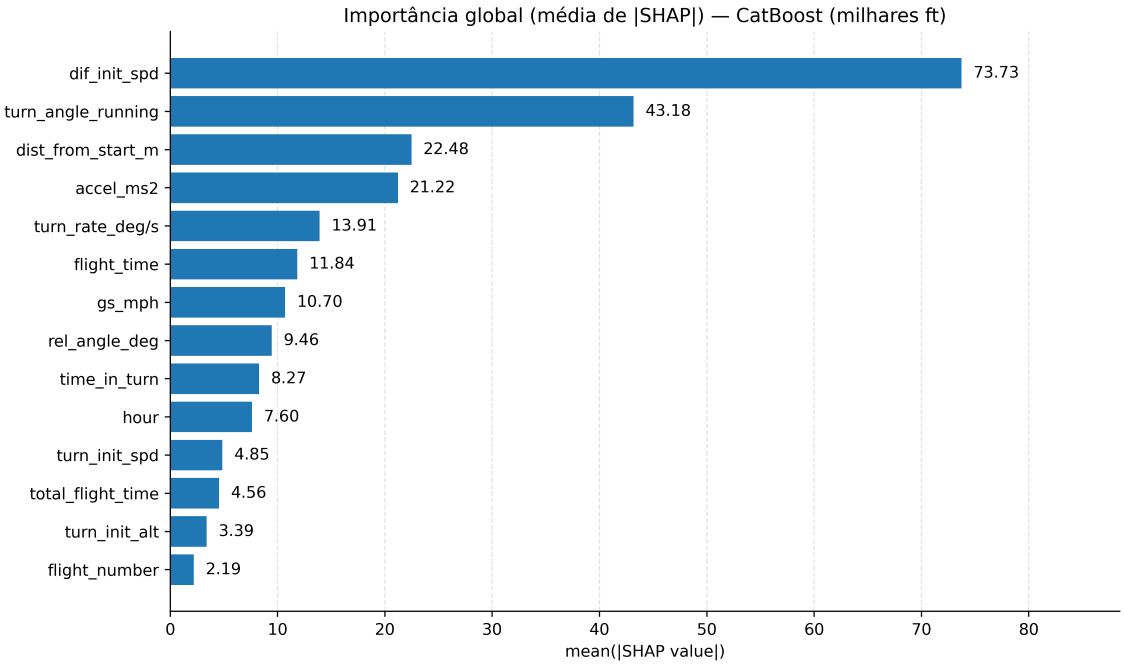

Figura 8 – Importância global das variáveis (média de |SHAP|) para o modelo CatBoost com `turn_init_alt` reparametrizado em milhares de pés.

Fonte: Autor.

A comparação entre os dois modelos sugere um compromisso claro entre precisão preditiva e robustez conceitual. Enquanto o modelo original apresenta melhor desempenho quantitativo no conjunto de teste, o modelo alternativo reduz a dependência de uma variável potencialmente associada a efeitos geográficos e operacionais específicos, favorecendo uma representação mais invariável da dinâmica da manobra.

Nesse sentido, o modelo com altitude reparametrizada não deve ser interpretado como estritamente inferior, mas como uma alternativa com maior potencial de generalização em cenários nos quais os dados não incluem faixas de pulverização contíguas, ou em aplicações envolvendo operadores distintos e regiões geográficas com altimetria diferente. Essa análise reforça a importância do uso conjunto de métricas de desempenho e ferramentas de explicabilidade na avaliação crítica de modelos preditivos aplicados a sistemas físicos complexos.

5.5 Limitações do modelo: análise dos piores casos

Com o objetivo de investigar limitações do modelo e as condições associadas aos maiores erros, analisaram-se as curvas cujo MAE por curva excede o percentil 95 (curvas $> P95$), em contraste com as demais curvas do conjunto de teste.

A explicabilidade foi avaliada por meio da razão entre a média de |SHAP| calculada

apenas nas curvas $> P95$ e a média de |SHAP| no conjunto de teste completo. A Figura 9 apresenta esse aumento relativo de magnitude das contribuições, destacando incrementos mais pronunciados para variáveis cinemáticas e de condição inicial da manobra, como `gs_mph` e `turn_init_spd`, seguidas por `accel_ms2`, `dist_from_start_m` e `dif_init_spd`. Ressalta-se que esse diagnóstico é descritivo, refletindo mudanças na magnitude média das contribuições do modelo nesse subconjunto, e não uma inferência causal.

Do ponto de vista dos dados, os histogramas comparativos (Apêndice B) indicam deslocamentos nas distribuições e maior incidência de caudas em variáveis que também exibem aumento relativo de |SHAP|, o que sugere que parte do erro elevado está associada a combinações de atributos menos representadas no conjunto de treinamento. Adicionalmente, observa-se concentração das curvas $> P95$ em faixas específicas de `turn_init_alt` e `hour`, bem como em valores mais altos de `dif_init_alt`, compatíveis com manobras de maior variação de altitude.

Em conjunto, os resultados apontam para limitações de generalização em sub-regiões do espaço de atributos associadas a condições operacionais específicas, potencialmente influenciadas por fatores não observados nas variáveis disponíveis (p. ex., diferenças de potência, técnica de pilotagem e condições meteorológicas), os quais podem alterar a dinâmica da curva de reversão e aumentar o erro preditivo.

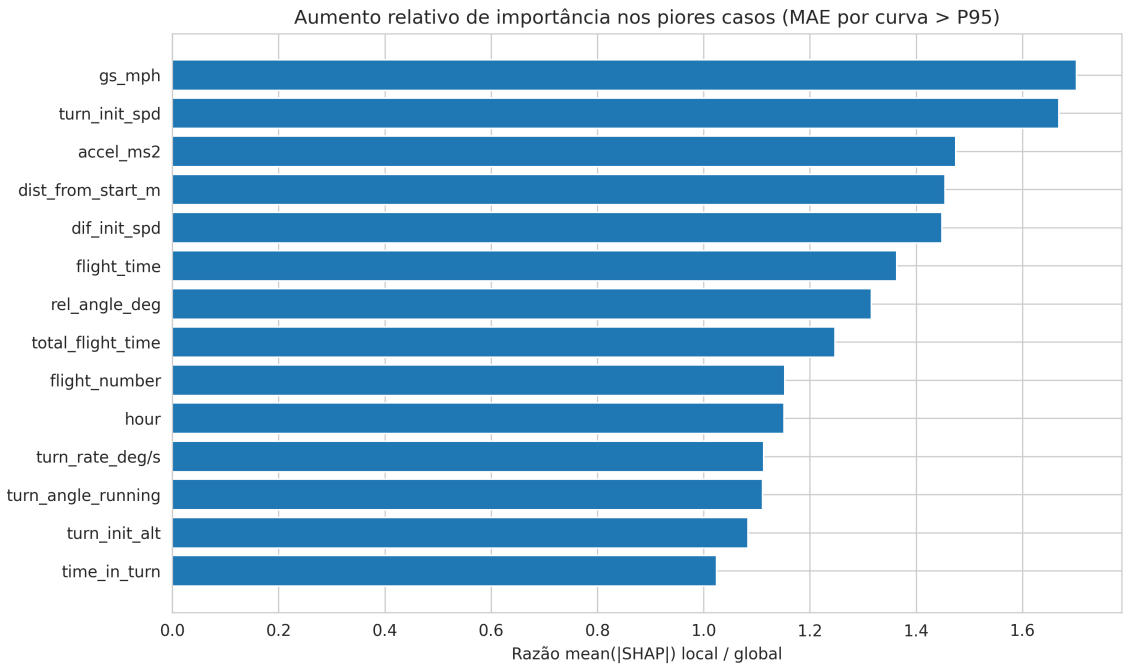

Figura 9 – Aumento relativo de importância das variáveis nos piores casos.

Fonte: Autor.

6 CONCLUSÕES

Este trabalho investigou a viabilidade de estimar a variação de altitude de aeronaves agrícolas durante curvas de reversão a partir de registros operacionais de sistemas AgNav–Guia, com foco na reconstrução da trajetória vertical em aeronaves equipadas com sistemas de navegação que não registram diretamente a altitude GPS. Para isso, foram coletados e preparados dados de voo, as curvas de reversão foram identificadas e segmentadas, e diferentes modelos de regressão supervisionada foram avaliados para a predição do Δalt, com posterior visualização do perfil vertical das manobras como evidência de viabilidade prática da abordagem.

No escopo experimental adotado, empregando registros provenientes de um único modelo de aeronave (Air Tractor 502A) e de um único operador, abrangendo um período superior a dois anos, o modelo CatBoost Regressor apresentou desempenho superior e foi selecionado como modelo final. Esse resultado foi sustentado por métricas globais e por métricas por curva, com erro médio absoluto (MAE) e percentil 95 do MAE por curva, $P95$(MAE), inferiores à escala típica da variável-alvo. A distribuição do MAE por curva indicou desempenho consistente no regime mais frequente de operação, embora com presença de cauda longa associada a um subconjunto reduzido de curvas com erro elevado.

A análise das curvas com MAE por curva acima do percentil 95 mostrou que os maiores erros tendem a emergir de combinações específicas de condições operacionais, e não de um único fator isolado. Os histogramas comparativos (Apêndice B) sugerem deslocamentos nas distribuições de variáveis-chave e concentração dessas curvas em faixas particulares de altitude inicial e de horário de operação, frequentemente associadas a manobras com maior variação de altitude (`dif_init_alt`). Esses achados são compatíveis com a hipótese de que variáveis não observadas nos registros disponíveis possam alterar a dinâmica da manobra e reduzir a acurácia do modelo em situações específicas.

Como contribuição principal, o modelo desenvolvido demonstra adequado potencial de representação da realidade, podendo ser aplicado para apoiar a reconstrução do perfil vertical de curvas de reversão quando a altitude não está disponível nos registros. Nesse contexto, a pesquisa possui relevância social ao contribuir para o aprimoramento de processos de investigação de ocorrências aeronáuticas, com perspectiva de apoiar a identificação de fatores contribuintes e, consequentemente, auxiliar a prevenção de novas ocorrências.

Como trabalho futuro, recomenda-se investigar modelos capazes de explorar dependências temporais ao longo da manobra, em especial redes neurais recorrentes do tipo *Long Short-Term Memory* (LSTM), com potencial de identificar padrões dinâmicos que

não são plenamente capturados pelos modelos de aprendizado de máquina utilizados neste estudo.

Recomenda-se também ampliar a base de dados para múltiplas aeronaves e operadores, permitindo avaliar robustez e generalização, incluindo uma análise comparativa por classes de aeronaves (menor versus maior porte/capacidade) a fim de identificar as condições em que os modelos apresentam uma melhor performance. Por fim, sugere-se o enriquecimento do conjunto de atributos com variáveis contextuais e proxies operacionais, incluindo integração com dados meteorológicos, parâmetros de motor oriundos de outros equipamentos embarcados e indicadores de qualidade do sinal (DGPS/registro), com o objetivo de reduzir a influência de variáveis latentes, melhorando a precisão dos modelos.

REFERÊNCIAS

APS TRAINING. **What is Angle of Attack? Three Critical Angles Explained**. 2025. [Internet]. Disponível em: https://info.apstraining.com/knowledge/what-is-angle-of-attack-three-critical-angles.

ARAÚJO, EDUARDO CORDEIRO DE. **DGPS: Aplicação Aérea de Precisão**. Pelotas, RS, 2005. Disponível em: https://sindag.org.br/wp-content/uploads/2021/07/DGPS_-Aplicacao-Aerea-de-Precisao.pdf. Acesso em: 18 jul. 2025.

CENTRO DE INVESTIGAÇÃO E PREVENÇÃO DE ACIDENTES AERONÁUTICOS (CENIPA). **Relatório Final A-046/CENIPA/2022: Acidente com a aeronave PR-AAM, modelo A188B, ocorrido em 14ABR2022**. Brasília, DF, 2024. Disponível em: https://sistema.cenipa.fab.mil.br/cenipa/paginas/relatorios/rf/pt/PR-AAM-14-04-2022-PUB..pdf. Acesso em: 18 jul. 2025.

CENTRO DE INVESTIGAÇÃO E PREVENÇÃO DE ACIDENTES AERONÁUTICOS (CENIPA). **Relatório Final A-133/CENIPA/2022: Acidente com a aeronave PP-OBL, modelo AT-502A, ocorrido em 18NOV2022**. Brasília, DF, 2024. Disponível em: https://sistema.cenipa.fab.mil.br/cenipa/paginas/relatorios/rf/pt/PP-OBL_Reabertura_PUB.pdf. Acesso em: 18 jul. 2025.

CENTRO DE INVESTIGAÇÃO E PREVENÇÃO DE ACIDENTES AERONÁUTICOS (CENIPA). **Investigação e Prevenção de Acidentes Aeronáuticos**. 2025. [Internet]. Disponível em: https://www2.fab.mil.br/cenipa/index.php/investigacoes.

CONFEDERAÇÃO DA AGRICULTURA E PECUÁRIA DO BRASIL (CNA). **Panorama do Agro**. 2025. [Internet]. Disponível em: https://www.cnabrasil.org.br/cna/panorama-do-agro.

FEDERAL AVIATION ADMINISTRATION (FAA). **Pilot's Handbook of Aeronautical Knowledge, Chapter 5: Aerodynamics of Flight**. [*S.l.*], 2024. Disponível em: https://www.faa.gov/sites/faa.gov/files/07_phak_ch5_0.pdf. Acesso em: 18 jul. 2025.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2006. Disponível em: https://link.springer.com/article/10.1007/s10994-006-6226-1.

GOMES, CLÁUDIO JÚNIOR OLIVEIRA. **Análise da Frota Aeroagrícola Brasileira de Aviões e Helicópteros 2024**. 1. ed. Brasília, DF, 2024. Disponível em: https://sindag.org.br/wp-content/uploads/2025/07/ANALISE-DA-FROTA-AEROAGRICOLA-BRASILEIRA-DE-AVIOES-E-HELICOPTEROS_2024_ver2.pdf. Acesso em: 18 jul. 2025.

HARRELL, F. E. **Regression Modeling Strategies**. 2. ed. [*S.l.: s.n.*]: Springer, 2015.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. Springer, 2009. Disponível em: https://hastie.su.domains/Papers/ESLII.pdf.

JASRA, S. e. a. **Literature Review of Machine Learning Techniques to Analyse Flight Data**. [*S.l.*], 2018. Disponível em: https://www.um.edu.mt/library/oar/bitstream/123456789/58839/3/AEGATS_Paper.pdf. Acesso em: 18 jul. 2025.

KE, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *In*: **Advances in Neural Information Processing Systems (NeurIPS)**. Long Beach, CA, USA: [*S.l.: s.n.*], 2017. Disponível em: https://proceedings.neurips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

LINDHOLM, A. *et al.* **Machine Learning: A First Course for Engineers and Scientists**. Cambridge University Press, 2022. Disponível em: http://smlbook.org.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, v. 30, p. 4765–4774, 2017.

MANGORTEY, E. e. a. Application of machine learning techniques to parameter selection for flight risk identification. *In*: **AIAA SciTech 2020 Forum**. [*S.l.: s.n.*], 2020. p. 1850. Disponível em: https://doi.org/10.2514/6.2020-1850. Acesso em: 18 jul. 2025.

MINISTÉRIO DA AGRICULTURA E PECUÁRIA (MAPA). **Serviços Aeroagrícolas**. 2025. [Internet]. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/aviacao-agricola/servicos-aeroagricolas.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. 7. ed. Hoboken, NJ: Wiley, 2018.

OZDEMIR, S.; SUSARLA, D. **Feature Engineering Made Easy**. Birmingham: Packt Publishing, 2018. ISBN 978-1-78728-760-0.

PROKHORENKOVA, L. *et al.* Catboost: Unbiased boosting with categorical features. *In*: **Advances in Neural Information Processing Systems (NeurIPS)**. Montréal, Canada: [*S.l.: s.n.*], 2018. Disponível em: https://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf.

PYCARET. **PyCaret Documentation**. 2023. Disponível em: https://pycaret.gitbook.io/docs/. Acesso em: 1 nov. 2025. [Internet]. Disponível em: https://pycaret.gitbook.io/docs/.

RASPET FLIGHT RESEARCH LABORATORY, MISSISSIPPI STATE UNIVERSITY. **Characterization of Agricultural Aircraft Performance Using Flight Log Data**. [*S.l.*], 2020. Disponível em: https://www.raspet.msstate.edu/sites/www.raspet.msstate.edu/files/2022-04/20200825%20Ag%20Data%20Model.pdf. Acesso em: 18 jul. 2025.

Scikit-learn Developers. **Model evaluation: quantifying the quality of predictions – Regression metrics**. 2025. https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics. Acesso em: 26 dez. 2025.

SHAP Developers. **SHAP Documentation**. 2024. https://shap.readthedocs.io. Disponível em: https://shap.readthedocs.io.

SINDICATO NACIONAL DAS EMPRESAS DE AVIAÇÃO AGRÍCOLA (SINDAG). **História**. 2025. [Internet]. Disponível em: https://sindag.org.br/historia/.

SISTEMA DE INVESTIGAÇÃO E PREVENÇÃO DE ACIDENTES AERONÁUTICOS (SIPAER). **Painel SIPAER**. 2025. [Internet]. Disponível em: https://painelsipaer.cenipa.fab.mil.br/extensions/Sipaer/Sipaer.html.

APÊNDICES

APÊNDICE A – ESTATÍSTICAS DESCRITIVAS DAS VARIÁVEIS DOS CONJUNTOS DE TREINAMENTO E TESTE

Tabela 9 – Estatísticas descritivas das variáveis do conjunto de teste

Estatística	gs_mph	accel_ms2	hour	turn_angle_running	turn_rate_deg/s	time_in_turn	turn_init_spd	dif_init_spd
count	18423	18423	18423	18423	18423	18423	18423	18423
mean	140.62	-0.05	14.28	-11.60	6.94	15.82	158.58	-17.96
std	17.52	1.55	3.28	119.84	6.43	11.09	8.55	18.95
min	76.60	-8.94	9.00	-270.00	0.00	0.00	132.50	-88.30
25%	131.20	-0.94	12.00	-120.00	1.00	7.00	152.80	-26.40
50%	141.80	0.04	13.00	-1.00	5.00	14.00	158.20	-14.80
75%	152.10	0.85	17.00	62.00	12.00	23.00	163.70	-5.10
max	202.60	12.07	21.00	216.00	35.00	60.00	193.00	45.50

Estatística	dist_from_start_m	flight_number	flight_time	total_flight_time	rel_angle_deg	turn_init_alt	dif_init_alt
count	18423	18423	18423	18423	18423	18423	18423
mean	590.13	2.18	1410.55	3754.42	-5.20	1366.81	176.20
std	363.20	1.62	914.74	2961.21	54.48	119.21	176.95
min	0.00	1.00	12.20	12.20	-179.42	1119.00	-161.00
25%	377.03	1.00	696.20	1365.20	-23.83	1250.00	53.00
50%	549.04	1.00	1279.00	2954.60	-0.38	1414.00	125.00
75%	752.68	3.00	1920.70	5630.10	11.45	1476.00	227.00
max	3006.64	9.00	4618.40	12092.60	179.97	1585.00	967.00

Fonte: Autor.

Tabela 10 – Estatísticas descritivas das variáveis do conjunto de treinamento

Estatística	gs_mph	accel_ms2	hour	turn_angle_running	turn_rate_deg/s	time_in_turn	turn_init_spd	dif_init_spd
count	73622	73622	73622	73622	73622	73622	73622	73622
mean	139.58	-0.05	14.18	-10.06	6.96	15.77	158.59	-19.01
std	18.19	1.61	3.16	118.61	6.40	10.98	8.63	19.69
min	78.10	-13.46	9.00	-222.00	0.00	0.00	125.10	-107.30
25%	129.80	-0.98	12.00	-113.00	1.00	7.00	152.80	-27.80
50%	141.00	0.04	13.00	0.00	5.00	14.00	157.80	-15.60
75%	151.70	0.89	17.00	61.00	12.00	23.00	164.20	-5.40
max	199.10	10.10	21.00	222.00	36.00	60.00	194.10	43.00

Estatística	dist_from_start_m	flight_number	flight_time	total_flight_time	rel_angle_deg	turn_init_alt	dif_init_alt
count	73622	73622	73622	73622	73622	73622	73622
mean	585.38	2.22	1341.17	3694.03	-2.80	1355.21	187.10
std	353.23	1.66	851.67	3022.40	50.87	122.57	181.85
min	0.00	1.00	2.80	2.80	-179.99	1106.00	-121.00
25%	374.18	1.00	669.40	1235.25	-21.05	1240.00	59.00
50%	544.91	1.00	1189.00	2694.50	-0.14	1358.00	131.00
75%	754.02	3.00	1861.20	5661.20	11.41	1467.00	243.00
max	3459.42	9.00	4707.40	12395.00	179.84	1608.00	928.00

Fonte: Autor.

APÊNDICE B – HISTOGRAMAS COMPARATIVOS DAS VARIÁVEIS: CURVAS $\leq$ P95 E $>$ P95

Os histogramas a seguir apresentam a comparação das distribuições relativas (em percentual) das variáveis entre os subconjuntos de curvas com erro por curva $\leq$ P95 e $>$ P95, utilizados na discussão das limitações do modelo.

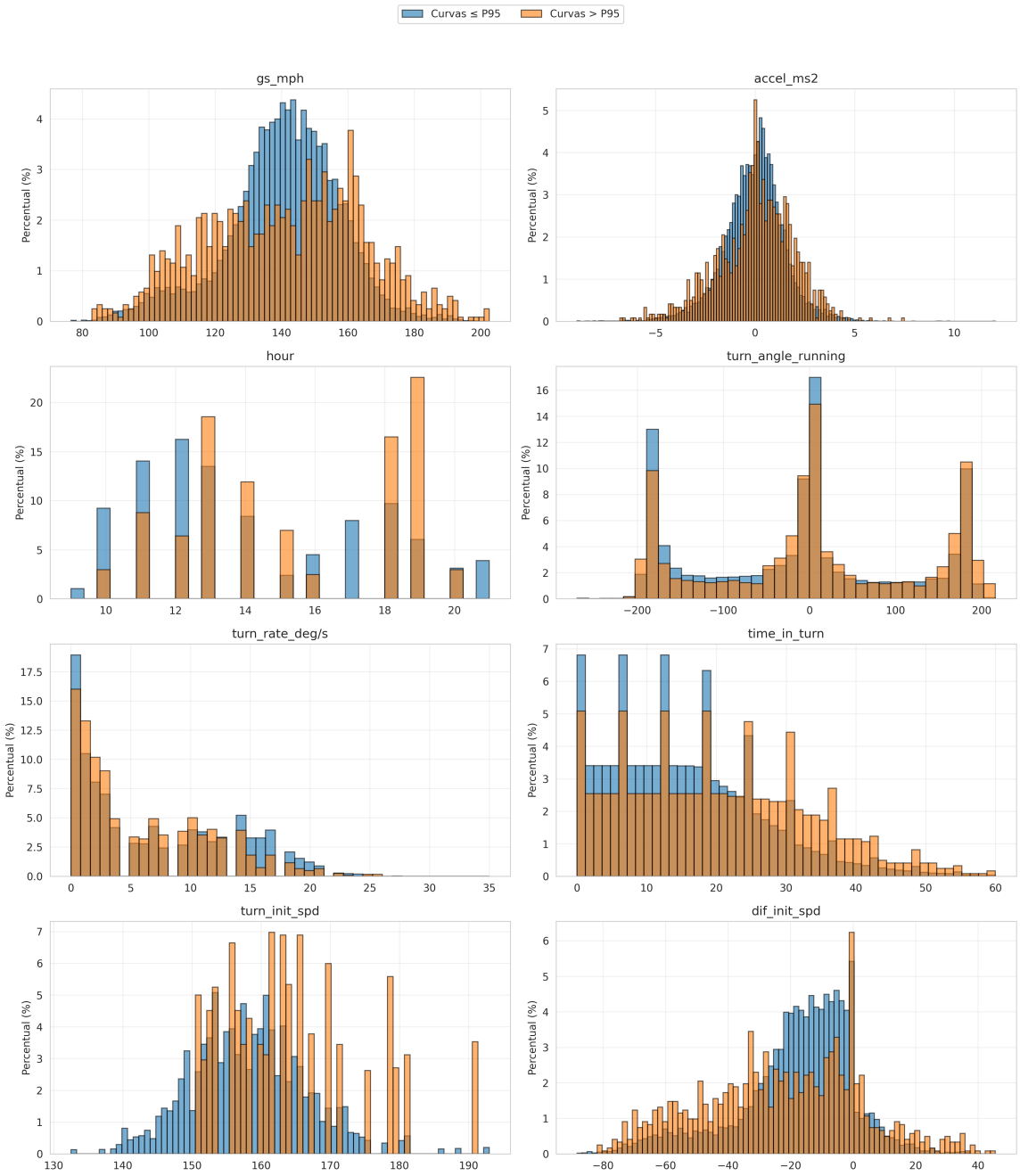

Figura 10 – Histogramas comparativos (percentual) das variáveis entre curvas $\leq$ P95 e $>$ P95 (parte A).

Fonte: Autor.

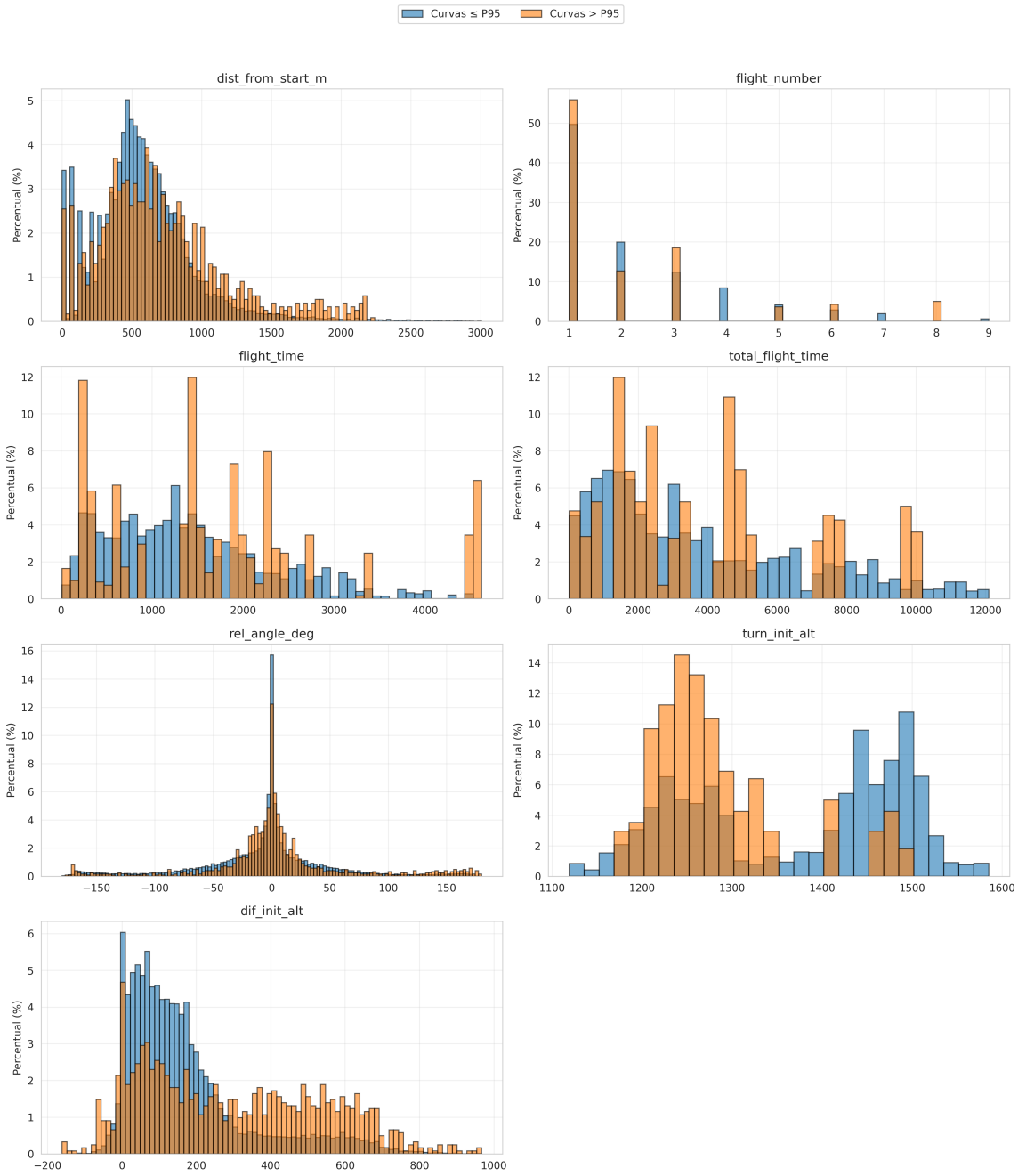

Figura 11 – Histogramas comparativos (percentual) das variáveis entre curvas $\leq$ P95 e $>$ P95 (parte B).

Fonte: Autor.