Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

INTRODUÇÃO À CIÊNCIA DE DADOS

Data: 21 de março de 2025

Relatório

Relatório produzido por Natália Reis, Felipe Tardieux e Moizes Filho com a finalidade de pôr em prática o conteúdo visto na disciplina de Introdução à Ciência de Dados.

1. Introdução

Faremos um estudo do preço das passagens aéreas na região metropolitana da Índia com o fim de entender quais são as variáveis que mais afetam o valor da passagem, e de que forma isso se dá, e desenvolver um modelo preditivo que preveja este valor baseado nos demais dados.

2. Fundamentos Teóricos e Metodológicos

Para atingirmos os objetivos destacados anteriormente, faremos a utilização de um conjunto de dados obtido no Kaggle [1] que será analisado posteriormente. Este conjunto de dados contém informações sobre opções de reserva de voos para as 6 principais cidades metropolitanas da Índia.

Os dados foram obtidos por web scraping no site EaseMyTrip, utilizando a ferramenta Octoparse. Eles foram coletados em duas partes: passagens de classe econômica e passagens de classe executiva, totalizando 300.261 opções de reservas de voos distintas e a coleta ocorreu ao longo de 50 dias, de 11 de fevereiro a 31 de março de 2022.

O conjunto de dados original foi levemente alterado com a remoção da coluna index e com a tradução das variáveis e de boa parte dos dados para o português para facilitar a visualização e o entendimento.

Dessa forma, o conjunto de dados utilizado é formado pelas seguintes variáveis:

Companhia Aérea: nome da companhia aérea; **Voo**: armazena informações sobre o código de voo do avião; **Cidade de Origem**: Cidade de onde o voo decola; **Horário de Partida**: característica categórica derivada, criada pela divisão de períodos de tempo em intervalos. Ela armazena informações sobre o horário de partida e possui 6 rótulos de tempo únicos; **Paradas**: número de paradas entre as cidades de origem e destino; **Horário de Chegada**: característica categórica derivada, criada pela divisão de intervalos de tempo em grupos. Possui seis rótulos de tempo distintos; **Cidade Destino**: cidade onde o voo irá pousar; **Classe**: classe do assento; possui dois valores distintos: executiva e econômica; **Duração (Horas)**: tempo de viagem, em horas; **Antecedência da Reserva (Dias)**: característica derivada que é calculada subtraindo a data da viagem pela data da reserva; **Preço (Rúpia Indiana)**: preço da passagem aérea.

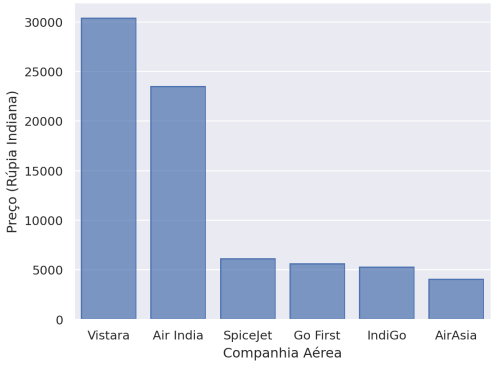
3. Aplicação

3.1 Análise Exploratória de Dados

Primeiro, analisemos de que forma a companharia aérea influência o preço da viagem. Ao comparar o valor da passagem por companhia, Gráfico 1, observamos que a Vistara e a Air India possuem um preço médio muito superior ao das demais companhias.

[1] https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

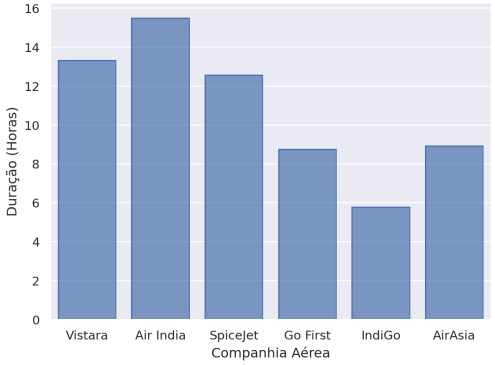
Gráfico 1: Preço Médio por Companhia Aérea



Fonte: Elaborado pelos autores, 2025.

A fim de encontrar uma justificativa para essa grande discrepância, analisemos o tempo médio de viagem e a proporção das classes aéreas para cada companhia.

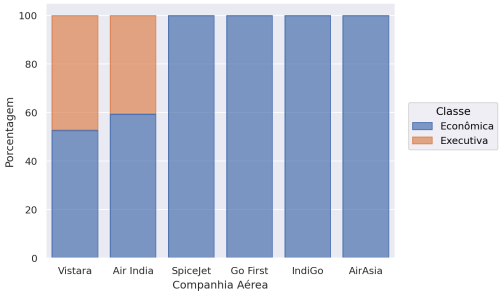
Gráfico 2: Tempo Médio de Viagem por Companhia Aérea



Fonte: Elaborado pelos autores, 2025.

Com o Gráfico 2 vemos que, de fato, as duas companhias que possuem o valor médio mais alto têm viagens mais longas em média do que as outras companhias. Porém, somente isto não chega a explicar a discrepância, pois a companhia SpiceJet, por exemplo, possui uma média de duração aproximada e, ainda assim, não possui passagens muito caras.

Gráfico 3: Proporção de Classe por Companhia Aérea



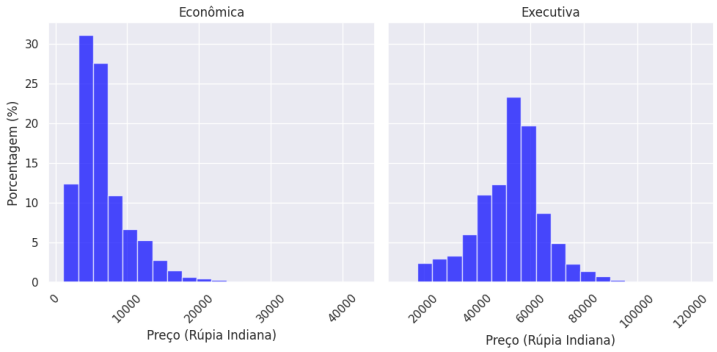
Fonte: Elaborado pelos autores, 2025.

No entanto, ao analizar a proporção por classe, vemos que a Vistara e a Air India são as únicas companhias que possuem viagens de classe executiva, o que parece explicar o preço médio tão alto.

Corroborando com isso, o Gráfico 4 nos mostra que as passagens da classe executiva são muito mais caras do que as passagens da classe econômica, com seu valor mínimo sendo quase o máximo da classe econômica. Isto pode ser visto claramente na distribuição do preço das passagens

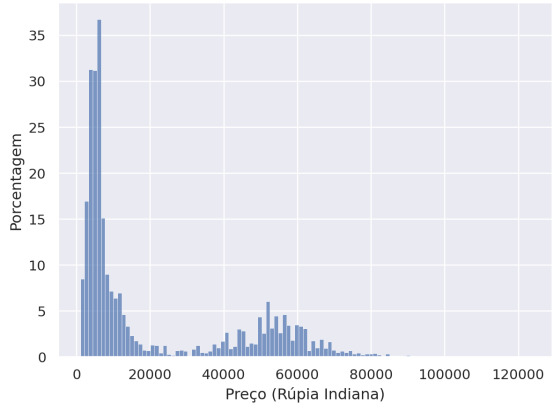
(Gráfico 5) que mostra a divisão clara que há entre dois focos de concentração dos preços que se dão devido às classes de voo. Portanto, a classe de voo é característica fundamental para prever o preço das passagens.

Gráfico 4: Distribuição do Preço da Passagem por Classe de Voo



Fonte: Elaborado pelos autores, 2025.

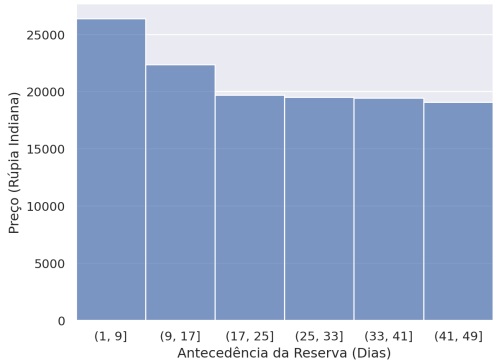
Gráfico 5: Distribuição do Preço das Passagens Aéreas



Fonte: Elaborado pelos autores, 2025.

Além disso, vejamos como a antecedência na compra da passagem afeta os preços. O Gráfico 6 nos mostra que comprar a passagem há poucos dias da viagem encarece a passagem, como esperado. No entanto, é interessante notar que após um certo ponto, cerca de 17 dias, aumentar a quantidade de dias antecipados influencia pouco o custo da passagem.

Gráfico 6: Preço Médio da Passagem por Tempo de Antecipação da Reserva



Fonte: Elaborado pelos autores, 2025.

3.2 Aprendizado de Máquina

Árvore de Regressão

Para realizar o treinamento, os dados foram divididos em 30% para teste e 70% para treino. Inicialmente, foram selecionadas aleatoriamente amostras de 100, 200 e 500 observações, respectivamente. Ao avaliar os modelos, percebeu-se que eles não estavam bem ajustados. Diante disso, foram realizadas duas novas tentativas. A primeira, com 10.000 observações aleatórias (doravante chamada de Treino 1), e a segunda, com todas as observações filtradas (sem outliers), doravante denominada Treino 2.

No Treino 1, foi utilizada uma árvore com profundidade máxima de 10 e validação cruzada com 10 dobras. Já no Treino 2, a profundidade máxima da árvore foi definida como 20, e foram utilizadas 20 dobras para a validação cruzada.

Conforme aumentavam-se a quantidade de observações notou-se redução no Mean Absolute Error (MAE) e no Mean Squared Error (MSE), tanto para o conjunto de teste quanto para as validações. No entanto, os valores dessas métricas permaneceram relativamente altos no Treino 1, em comparação com os valores reais das passagens.

Em termos de desempenho, as validações cruzadas do Treino 1 foi capaz de explicar, em média, 94,5% da variância nos preços (R^2). O Root Mean Squared Logarithmic Error (RMSLE) médio foi de 0,2919, indicando um baixo erro logarítmico. Além disso, o Mean Absolute Percentage Error (MAPE) médio atingiu 22,43%. Ao fazer as predições, os índices se aproximam dos encontrados nas validações com ligeiramente maior e RMSLE e MAPE ligeiramente menores, o que indica boa capacidade de generalização do modelo.

O Treino 2 conseguiu prever os dados do teste com MAE de 1158.2872 o que é uma grande diferença considerando o Treino 1. O RMSE também teve diminuição e o R^2 obteve um ligeiro aumento de 0.01. Uma grande diferença pode ser vista quando compara-se o RMSLE, que de 0.2721 para 0.1457 e o MAPE de 20.87% foi para 6.86%. O Treino 2 também manteve a característica de boa generalização, obtendo indicadores melhores no teste que na média das validações cruzadas, portanto considera-se que o modelo está bem ajustado.

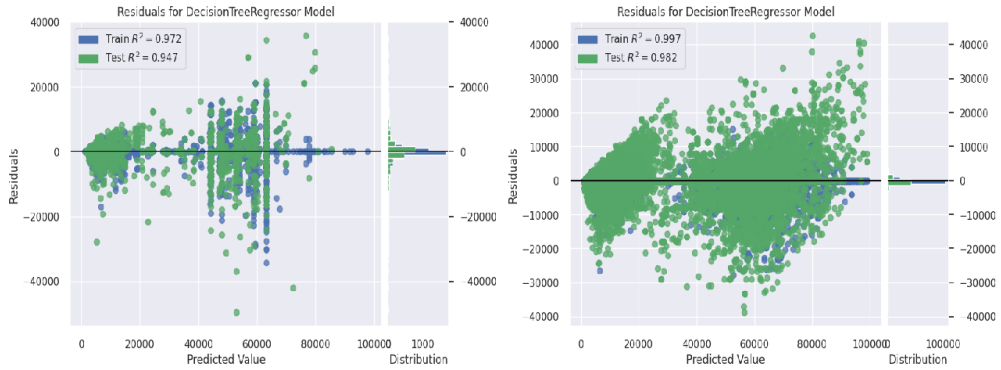
Tabela 1: Métricas para Avaliação do Treino 1 e Treino 2

Métrica	MAE	MSE	RMSE	R2	RMSLE	MAPE
Média das Validações do Teste Treino 1	3012.3549	28321694	5312.9808	0.9452	0.2919	0.2243
Validação do Teste Treino 1	2902.6059	27994629	5290.9952	0.9469	0.2781	0.2087
Média das Validações do Treino Treino 2	1158.2872	8972446	2993.7364	0.9825	0.1755	0.0882
Validação do Teste Treino 2	883.4082	7148779	2673.7202	0.9860	0.1457	0.0686

Fonte: Elaborado pelos autores, 2025.

No Gráfico 7 observa-se que a medida que os valores das passagens aumentam o modelo fica mais instável com resíduos maiores, enquanto nos valores mais baixos os resíduos variam menos. Quanto a distribuição dos resíduos, ela se assemelha a uma normal, para ambos treinos. O coeficiente de determinação (R^2) está acima de 0.9 no Treino 1 e no Treino 2 está acima de 0.98.

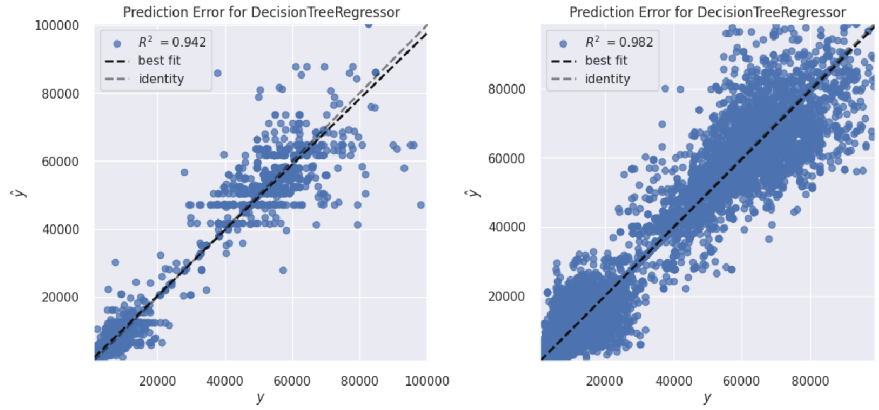
Gráfico 7: Resíduos para o Modelo de Árvore de Regressão



Fonte: Elaborado pelos autores, 2025.

O Gráfico 8: de Erro de Predição reforça o padrão de um melhor ajuste para passagens com valores menores enquanto as de valores maiores possuem maior variância nas predições, cometendo erros maiores principalmente no Treino 1. Os pontos se espalham de forma aleatória em torno da reta de referência e a reta de identidade se aproxima muito a reta do melhor ajuste, principalmente no Treino 2.

Gráfico 8: Erro de Predição



Fonte: Elaborado pelos autores, 2025.

Ao analisar a importância das variáveis destaca-se, em ambos os treinos, a Classe da passagem que possui um impacto de mais de 90% nas previsões do modelo, em seguida estão as variáveis Duração (Horas) e Voo, ambas com menos de 10% de impacto.

Ao avaliar o Treino 1 otimizado, conclui-se que ele segue os mesmos índices do modelo não otimizado. Portanto o Treino 1 não otimizado conseguiu captar bem todas as características dos dados e possui boa capacidade de generalização, sendo descartável o modelo otimizado visando redução de custo computacional. O Treino 2 não foi otimizado, uma vez que a otimização com mais de 300 mil observações demoraria muito e como visto, o Treino 1 otimizado não se mostrou superior ao não-otimizado.

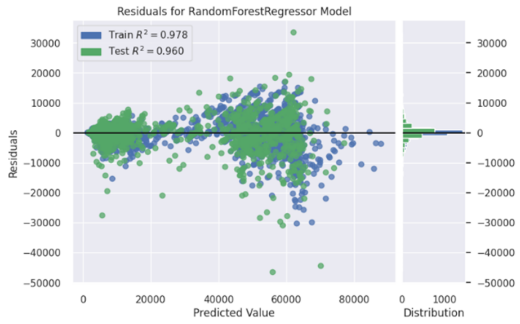
Floresta Aleatória para Regressão

Realizamos o modelo com um total de 10000 observações onde 70% foram usadas para treinamento e 30% para teste. Na configuração do modelo, selecionamos uma semente para reprodutibilidade e o número de árvores de decisão na floresta (n_estimators) foi de 100 com um limite máximo de profundidade de 10 e 10 partes (folds).

Com essa aplicação temos que em média o modelo apresentou uma variação de 96% nos preços e com uma média de erro absoluto de percentual (MAPE) de 19,77% no treinamento.

No Gráfico de Resíduos para Floresta de Regressão do modelo observou-se que à medida que há um aumento nos valores das passagens, há uma maior instabilidade para os resíduos maiores, já para os valores baixos os resíduos variam menos. O coeficiente de determinação (R^2) para os dados do teste é maior que 0,9, indicando que se tem um ótimo ajuste.

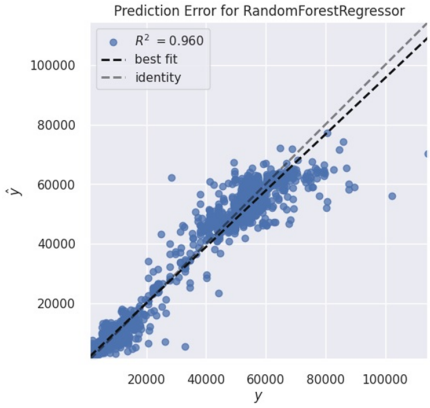
Gráfico 9: Resíduos para Floresta de Regressão



Fonte: Elaborado pelos autores, 2025.

O Gráfico de Erro mostra que não há uma grande variação para as passagens de menor valor, enquanto as de valores maiores possuem uma variância nas predições mesmo que sutil em relação às de menor valor, causando assim um maior número de erros.

Gráfico 10: Gráfico de Erro de Predição



Fonte: Elaborado pelos autores, 2025.

Na realização do gráfico da importância das variáveis destaca-se que a variável Classe possui um impacto de mais de 90% em relação aos das previsões do modelo, e as variáveis Duração e Voo não possuem um grande impacto.

Fazendo a predição do modelo se tem que o R^2 não teve um aumento significativo. RMSLE e MAPE tem uma diminuição pequena, indicando que mesmo com uma pequena alteração nos valores há uma boa capacidade de generalização do modelo.

Fazendo a otimização do modelo há uma piora quando comparando os indicadores com os que não são otimizados em que R^2 diminui enquanto RMSLE e MAPE aumentam. Visualizando os gráficos de resíduos e de erro, vemos que também não há mudança significativa. Também não havendo mudança quanto à importância das variáveis seguindo o mesmo padrão do modelo não otimizado. E as predições do modelo segue o mesmo nível de acurácia dos não otimizados.

Concluímos que o modelo não otimizado capta bem as características dos dados e possui boa capacidade de generalização, sendo descartável o modelo otimizado visando redução de custo computacional.

4. Conclusão

A análise realizada neste estudo permitiu identificar os principais fatores que influenciam o preço das passagens aéreas nas seis maiores cidades metropolitanas da Índia, com base no conjunto de dados examinado. Dentre as variáveis analisadas, a classe do voo se destacou como o fator de maior impacto na predição dos preços, sendo responsável por mais de 90% da variabilidade observada nos modelos de aprendizado de máquina. Esse resultado é corroborado pela análise exploratória, que demonstrou uma separação clara nos preços entre as classes, com a classe executiva apresentando valores significativamente maiores.

Além disso, foi observado que o aumento no número de observações utilizadas no treinamento dos modelos resultou em uma melhoria substancial na qualidade das predições. Esse aprimoramento indica que a quantidade de dados é um elemento crucial para o ajuste e a generalização dos modelos preditivos neste contexto.

Por fim, a comparação entre os modelos otimizados e não otimizados revelou que a otimização não trouxe ganhos significativos em desempenho ao mesmo tempo em que demandaria maior custo computacional. Tanto na árvore de regressão quanto na floresta aleatória, os modelos não otimizados demonstraram captar adequadamente as características dos dados e exibiram boa capacidade de generalização, tornando a otimização dispensável.

5. Contribuições da equipe

Natália Viviane Silva Reis: *Aplicação de modelo de aprendizado de máquina* – 50%; *Relatório (descrição e análise)* – 33%

Felipe Maia Tardieux: *Aplicação de modelo de aprendizado de máquina* – 50%; *Análise exploratória de dados* – 30%; *Relatório (descrição e análise)* – 33%

Moizes Claudino Bezerra Filho: *Análise exploratória de dados* – 70%; *Relatório (descrição e análise)* – 33%

6. Referências

https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction