

Leonardo Pereira Gonçalves, Henrique Murakami Silva, Leonardo Santos Pereira, Kayque Soares, Luis Fernando Silva Ferreira

Análise Preditiva de Preços de Smartphones para o Mercado de 2025

UEMG - Sistemas de informação
Mineração de Dados - Rení Aparecido Norberto Pinto

Passos - MG

15 de novembro de 2025

Resumo

Este trabalho apresenta o desenvolvimento de um projeto de mineração de dados voltado para a previsão de preços de smartphones, utilizando um conjunto de dados simulado para o mercado de 2025. O estudo percorre todas as etapas essenciais, desde a organização do ambiente colaborativo de trabalho e a carga inicial das informações, até a verificação preliminar da integridade e consistência dos dados. A extração e a primeira inspeção foram realizadas no Google Colab, com o apoio da biblioteca Pandas, permitindo identificar as transformações e limpezas necessárias para preparar o dataset para análises mais profundas. A partir desse processo inicial, o projeto evolui para a construção e avaliação de modelos de regressão, cujo propósito é estimar o preço final de um smartphone com base em suas especificações técnicas.

Palavras-chave: mineração de dados, aprendizado de máquina, previsão de preços, regressão, análise de dados.

Sumário

1. Introdução.....	4
1.1 Contextualização do Problema	4
1.2 Problema de Pesquisa	4
1.3 Justificativa.....	4
1.4 Objetivo Geral	4
1.5 Objetivos Específicos.....	4
1.6 Organização do Relatório	5
2. Referencial Técnico e Metodológico	5
2.1 Mineração de Dados e Aprendizado de Máquina	5
2.2 Modelos de Regressão	6
2.2.1 Regressão Linear	6
2.2.2 Random Forest Regressor	6
2.3 Processo de ETL	6
2.4 Ferramentas e Ambiente de Desenvolvimento.....	7
3. Metodologia Aplicada ao Projeto	7
3.1 Descrição do Dataset.....	7
3.2 Processo ETL.....	8
3.2.1 Extração e Carga Inicial.....	8
3.2.2 Análise Preliminar e Direcionamento da Limpeza.....	8
3.2.3 Implementação da Etapa de ETL	9
3.2.3.1 Transformações Aplicadas.....	10
3.2.3.2 Código Utilizado.....	11
3.2.4 Resultado Final	11
3.3 Estrutura dos Notebooks.....	11
3.4 Pipeline Analítico	12
4. Análise Exploratório dos Dados	13
5. Modelagem e Avaliação dos Resultados.....	14
6. Conclusão.....	15
7. Referências	16

1. Introdução

1.1 Contextualização do Problema

O preço de um smartphone decorre da combinação de vários componentes que influenciam tanto o desempenho quanto o posicionamento do aparelho no mercado. Recursos como maior quantidade de RAM, processadores mais potentes e câmeras de qualidade superior geralmente resultam em dispositivos mais caros. No entanto, apenas observar as especificações não é suficiente para determinar a participação exata de cada variável na formação do preço. Por isso, analisar essas informações com técnicas de ciência de dados permite transformar os dados brutos em conhecimento mais objetivo e fundamentado.

1.2 Problema de Pesquisa

Quais características técnicas são determinantes para o preço de um smartphone no mercado global de 2025?

1.3 Justificativa

A compreensão dos fatores que influenciam o preço dos smartphones fornece subsídios importantes para analistas, consumidores, equipes de marketing e fabricantes,

permitindo melhor entendimento do posicionamento de produtos e dos elementos que agregam maior valor ao consumidor final.

1.4 Objetivo Geral

Desenvolver um modelo de aprendizado de máquina capaz de prever o preço de smartphones com base em suas especificações técnicas.

1.5 Objetivos Específicos

- Realizar análise exploratória dos dados.
- Avaliar o posicionamento de marcas e variações de preço.
- Treinar e comparar modelos de regressão.

- Identificar as variáveis com maior influência sobre o preço final.
- Gerar um modelo preditivo com bom desempenho.

1.6 Organização do Relatório

O relatório é estruturado em seis capítulos. O primeiro apresenta o contexto, o problema e os objetivos. O segundo descreve o referencial técnico utilizado ao longo do estudo. O terceiro detalha a metodologia aplicada, com foco no processo de preparação dos dados. O quarto capítulo analisa o comportamento do conjunto de dados por meio da exploração gráfica e estatística. O quinto apresenta os modelos aplicados e discute os resultados obtidos. O sexto e último capítulo reúne as conclusões gerais e as sugestões de continuidade para trabalhos futuros.

2. Referencial Técnico e Metodológico

O desenvolvimento deste projeto se apoia em princípios de mineração de dados, aprendizado de máquina e processos de preparação de dados amplamente utilizados em estudos analíticos. Essas bases teóricas fornecem sustentação para compreender as etapas do projeto, bem como a lógica por trás das técnicas aplicadas na análise exploratória e na modelagem preditiva.

2.1 Mineração de Dados e Aprendizado de Máquina

A mineração de dados é o processo de examinar grandes conjuntos de informações para identificar padrões, relações e estruturas que não são facilmente percebidas de forma direta. Essa área se relaciona de maneira direta com o aprendizado de máquina, que consiste em técnicas capazes de aprender automaticamente a partir de dados, construindo modelos que realizam previsões ou classificações com base em exemplos anteriores.

No contexto deste projeto, o foco está no aprendizado supervisionado, especificamente na tarefa de regressão, que busca prever valores numéricos contínuos. A partir de um conjunto de exemplos composto por diversas características técnicas e o preço correspondente de cada smartphone, o modelo tenta identificar uma relação matemática que permita estimar o preço de novos aparelhos com base em suas especificações.

2.2 Modelos de Regressão

Modelos de regressão são utilizados quando o objetivo é prever valores contínuos. Eles buscam compreender como uma ou várias variáveis independentes influenciam uma variável dependente. Neste projeto, essas variáveis independentes incluem atributos como memória RAM, armazenamento, processador e câmeras, enquanto a variável dependente é o preço final dos smartphones.

Dois modelos foram selecionados por sua relevância, simplicidade e capacidade de capturar padrões distintos: a Regressão Linear e o Random Forest Regressor. Esses modelos apresentam abordagens conceitualmente diferentes, permitindo uma comparação mais clara entre métodos tradicionais e algoritmos mais avançados.

2.2.1 Regressão Linear

A Regressão Linear é um método estatístico clássico que busca modelar a relação entre variáveis por meio de uma combinação linear. Ela parte do princípio de que o preço de um smartphone pode ser explicado como a soma ponderada de suas características técnicas. Embora simples, esse modelo serve como ponto de partida importante, pois oferece facilidade de interpretação e fornece um parâmetro de comparação para modelos mais sofisticados.

2.2.2 Random Forest Regressor

O Random Forest é um algoritmo de aprendizado de máquina baseado no uso de múltiplas árvores de decisão. Ele funciona criando diversas árvores, cada uma analisando diferentes subconjuntos dos dados e das variáveis. O resultado final é obtido a partir da média das previsões feitas por todas essas árvores. Esse processo torna o modelo mais robusto, capaz de capturar relações não lineares e comportamentos complexos nos dados, o que é muito útil em cenários onde múltiplas variáveis influenciam a saída de maneiras diferentes.

2.3 Processo de ETL

O processo de ETL, sigla para Extração, Transformação e Carga, é essencial para garantir que o conjunto de dados esteja adequado às etapas de análise e modelagem. Ele envolve a leitura dos dados brutos, a aplicação de transformações necessárias para corrigir problemas e padronizar informações e, por fim, a criação de um arquivo final limpo e pronto para ser utilizado.

Neste projeto, o ETL contemplou a leitura do arquivo original, a verificação inicial da estrutura dos dados, a limpeza de inconsistências, a conversão de variáveis categóricas e a aplicação de técnicas de codificação para transformar informações textuais em dados numéricos compatíveis com modelos matemáticos.

2.4 Ferramentas e Ambiente de Desenvolvimento

O estudo foi conduzido utilizando linguagem Python no ambiente Google Colab. Essa plataforma permitiu a execução dos notebooks de forma simples, integrada e colaborativa. As bibliotecas Pandas, NumPy e Matplotlib foram fundamentais para manipulação, análise e visualização dos dados, enquanto a biblioteca Scikit-Learn forneceu os métodos necessários para construção dos modelos de regressão. Todo o processo foi versionado e organizado por meio de um repositório no GitHub, garantindo controle e rastreabilidade das etapas desenvolvidas.

3. Metodologia Aplicada ao Projeto

A metodologia utilizada neste projeto foi estruturada de forma sequencial, passando pelas etapas de descrição do conjunto de dados, preparação por meio do processo de ETL, análise exploratória e modelagem preditiva. Cada uma dessas etapas é fundamental para garantir que os resultados produzidos sejam consistentes, coerentes e representem adequadamente o comportamento dos dados disponíveis.

3.1 Descrição do Dataset

O conjunto de dados utilizado no projeto é denominado *Global Mobile Prices 2025 Extended* e contém informações simuladas de mil smartphones previstos para o mercado de 2025. O arquivo disponibiliza quinze variáveis relacionadas às especificações técnicas dos aparelhos, incluindo memória RAM, armazenamento interno, processador, sistema operacional, câmeras, capacidade de bateria, suporte a 5G, marca, mês de lançamento e preço final em dólares.

A inspeção inicial mostrou que não havia valores nulos e que todas as variáveis estavam registradas de forma consistente, o que facilitou o processo de preparação dos dados.

3.2 Processo ETL

O processo de ETL teve como objetivo preparar o dataset original para análises mais aprofundadas e para aplicação dos modelos de regressão. Esse processo envolveu a leitura do dataset bruto, a verificação das suas características iniciais e a aplicação das transformações necessárias para torná-lo adequado às etapas seguintes.

3.2.1 Extração e Carga Inicial

A primeira etapa consistiu na importação do arquivo CSV para o ambiente Python utilizando a biblioteca Pandas. Esse procedimento viabilizou a visualização inicial da estrutura do dataset, incluindo número de registros, nomes das colunas e tipos das variáveis.

A seguir está o código utilizado para a extração e carga inicial:

```
# Importando a biblioteca Pandas
import pandas as pd

# Caminho do arquivo no ambiente Colab
caminho = '/content/Global_Mobile_Prices_2025_Extended.csv'

# Carga do CSV para um DataFrame
df = pd.read_csv(caminho)

# Exibindo informacoes gerais (tipos de dados e contagem de nulos)
df.info()
```

Listing 1 – Código para carga e inspeção inicial dos dados.

3.2.2 Análise Preliminar e Direcionamento da Limpeza

A execução do código utilizado na carga inicial permitiu verificar que o conjunto de dados é composto por mil registros distribuídos em quinze colunas, e que não havia qualquer valor ausente. Essa característica tornou o processo de preparação mais simples, já que não foi necessário aplicar técnicas de imputação ou remoção de linhas incompletas. No entanto, a análise visual das

primeiras linhas revelou a necessidade de algumas correções para garantir a qualidade do dataset antes das etapas de análise e modelagem.

Durante a inspeção, observou-se que a coluna *model*, responsável por identificar o nome de cada aparelho, apresentava ruídos numéricos ao final de alguns registros, como em “A98 111”. Esses elementos não possuem relevância para a análise e poderiam comprometer a padronização dos modelos, tornando necessária a remoção desses números adicionais. Além disso, identificou-se que a coluna *5g_support* estava armazenada em formato textual, com valores como “Yes” e “No”, e precisava ser convertida para valores numéricos binários, como 1 e 0, a fim de possibilitar seu uso adequado em modelos matemáticos. Outras colunas categóricas, como *processor*, também exigiam tratamento adicional, como a aplicação de técnicas de codificação para representar suas categorias de forma numérica, sendo o One-Hot Encoding o método apropriado para esse caso.

3.2.3 Implementação da Etapa de ETL

Após a análise preliminar apresentada anteriormente, foi desenvolvido o processo completo de ETL (Extração, Transformação e Carga), responsável por limpar, padronizar e preparar o conjunto de dados para a análise exploratória e para a posterior modelagem preditiva. O processo foi implementado utilizando Python e a biblioteca Pandas no ambiente Google Colab, garantindo um fluxo simples e eficiente. Conforme explicado por Tecmundo (2025), o processo de ETL é composto pelas etapas de extração dos dados brutos, transformação para correção e padronização das informações e, por fim, a carga do conjunto de dados já tratado para uso nas próximas etapas do projeto.

```
#Importando as bibliotecas essenciais import
pandas as pd

#1. Extrair

caminho_do_arquivo = '/content/Global_Mobile_Prices_2025_Extended.csv' df =
pd.read_csv(caminho_do_arquivo)

#2. Transformar

#Limpeza da coluna model (remover valores irrelevantes) df['
model'] = df['model'].str.replace(r'\s\d+$', '', regex=True)
```

```

#Converter a coluna 5g_support para valores numericos

df['5g_support'] = df['5g_support'].apply(lambda x: 1 if x == 'Yes'
                                           else 0)

#One-Hot Encoding para variaveis categoricas

colunas_categoricas = ['brand', 'os', 'processor', 'release_month']

df = pd.get_dummies(df, columns=colunas_categoricas, drop_first=True)

```

Listing 2 – Código completo de ETL utilizado para limpeza e transformação dos dados.

3.2.3.1 Transformações Aplicadas

Com base nos problemas identificados durante a análise preliminar do conjunto de dados, foi necessário aplicar um conjunto de transformações para garantir a padronização das variáveis e a adequação do dataset às etapas analíticas seguintes. A primeira transformação realizada foi a limpeza da coluna *model*, que apresentava ruídos numéricos ao final de alguns nomes de aparelhos. Esses valores adicionais não possuíam relevância e poderiam gerar inconsistências durante as análises, motivo pelo qual foram removidos para garantir a uniformidade da informação.

Além disso, a variável *5g_support* encontrava-se registrada como texto, com valores como “Yes” e “No”, exigindo conversão para valores numéricos binários. Essa transformação foi necessária, pois modelos matemáticos e algoritmos de aprendizado de máquina dependem de dados numéricos para processar e aprender padrões de forma eficiente. Outras colunas categóricas do dataset, como *brand*, *os*, *processor* e *release_month*, também foram convertidas para variáveis numéricas por meio da técnica de One-Hot Encoding. Essa abordagem consiste em transformar cada categoria em uma nova coluna, representando sua presença ou ausência, garantindo que nenhuma informação importante fosse perdida e que todas as variáveis estivessem compatíveis com os modelos preditivos.

Essas transformações foram essenciais para a criação de um conjunto de dados limpo e devidamente estruturado, permitindo que os próximos passos de análise exploratória e modelagem fossem realizados de forma eficiente e sem incompatibilidades técnicas.

3.2.3.2 Código Utilizado

O processo de transformação foi implementado em Python utilizando a biblioteca Pandas, de modo a garantir que todas as variáveis estivessem devidamente organizadas para as etapas posteriores. O código apresentado a seguir reúne todas as operações realizadas durante o ETL, desde a leitura do arquivo original até a geração do dataset final em sua forma limpa e padronizada. Ele inclui a remoção dos ruídos presentes na coluna *model*, a conversão da variável *5g_support* para valores binários e a aplicação de One-Hot Encoding nas colunas categóricas, resultando em um arquivo final preparado para análises estatísticas e aplicação dos modelos preditivos.

3.2.4 Resultado Final

Ao final do processo de ETL, foi gerado o arquivo *dados_limpos_para_analise.csv*, que reúne o conjunto de dados totalmente tratado e pronto para ser utilizado nas etapas seguintes do projeto. Esse arquivo inclui todas as variáveis numéricas e categóricas devidamente convertidas, além da remoção de inconsistências identificadas na análise inicial. As transformações aplicadas garantiram que o dataset estivesse padronizado, organizado e livre de ruídos, o que contribuiu diretamente para a qualidade das análises exploratórias e dos modelos de regressão utilizados posteriormente.

O resultado dessa etapa representa a consolidação de um dataset robusto, completamente compatível com métodos estatísticos e algoritmos de aprendizado de máquina. Assim, todas as variáveis foram transformadas para um formato adequado, sem valores ausentes e sem informações redundantes. Com isso, tornou-se possível avançar para a análise exploratória com segurança, assegurando que os padrões encontrados nos dados refletissem de forma fiel a estrutura e o comportamento do conjunto original.

3.3 Estrutura dos Notebooks

O projeto foi organizado em quatro notebooks principais, cada um responsável por uma etapa específica do fluxo de trabalho em ciência de dados. Essa separação foi importante para manter o processo estruturado, facilitar a colaboração entre os integrantes e permitir que cada etapa pudesse ser revisada e executada de forma independente. A divisão dos notebooks segue a ordem

lógica do pipeline, começando pela preparação dos dados, avançando pela análise exploratória e finalizando com a modelagem e avaliação dos resultados.

O primeiro notebook contém toda a implementação do processo de ETL, reunindo as etapas de extração, transformação e carga do dataset. Nele foram realizadas as correções e padronizações necessárias para garantir que todas as variáveis estivessem adequadas ao uso posterior. O segundo notebook é dedicado exclusivamente à análise exploratória dos dados e apresenta gráficos, distribuições e relações entre variáveis. Essa etapa permitiu compreender melhor a estrutura do conjunto de informações e identificar padrões relevantes. O terceiro notebook concentra a aplicação dos modelos preditivos, incluindo a separação entre conjuntos de treino e teste, o treinamento dos algoritmos e o cálculo das métricas de desempenho. Por fim, o quarto notebook reúne a avaliação dos resultados obtidos, apresentando a interpretação das métricas, as justificativas para seleção do modelo final e uma síntese dos achados obtidos ao longo do processo.

Essa organização modular permitiu que o projeto fosse conduzido de forma clara, sequencial e rastreável, garantindo que cada etapa pudesse ser consultada quando necessário e facilitando o entendimento geral de todo o fluxo analítico.

3.4 Pipeline Analítico

O pipeline analítico seguido neste projeto foi desenvolvido de forma sequencial e lógica, respeitando as etapas fundamentais da mineração de dados. Ele foi composto por um fluxo contínuo que começa com a preparação do conjunto de dados e evolui para a análise e modelagem, permitindo que cada fase se apoie nos resultados da anterior. Esse fluxo garantiu que todo o processo fosse conduzido com organização e coerência, desde o momento em que o dataset foi carregado até a interpretação dos resultados produzidos pelos modelos.

A primeira etapa do pipeline foi o processo de ETL, que teve a função de preparar o conjunto de dados, corrigindo inconsistências e transformando variáveis para que estivessem adequadas ao uso em análises estatísticas e algoritmos de aprendizado de máquina. Em seguida, realizou-se a análise exploratória, que buscou compreender o comportamento das variáveis, identificar tendências e avaliar a relação entre características técnicas e o preço dos smartphones. A partir dessas observações, tornou-se possível definir

expectativas mais claras sobre o desempenho dos modelos e selecionar os algoritmos adequados para o problema.

Com o dataset limpo e a análise inicial concluída, avançou-se para a etapa de modelagem preditiva, onde os modelos de regressão foram treinados, testados e avaliados. Essa fase permitiu observar qual algoritmo melhor representava as relações entre as variáveis e oferecia maior precisão nas previsões. Por fim, a última etapa do pipeline consistiu na avaliação e interpretação dos resultados obtidos, encerrando o ciclo analítico com a seleção do modelo final e a análise das conclusões gerais do estudo.

4. Análise Exploratório dos Dados

A análise exploratória dos dados teve como principal objetivo compreender o comportamento geral do conjunto de informações após as etapas de limpeza e transformação realizadas no processo de ETL. Esse momento foi essencial para identificar padrões, relações entre as variáveis e possíveis tendências que pudessem influenciar tanto a modelagem quanto a interpretação dos resultados. Com o dataset já padronizado, tornou-se possível observar com clareza como as características técnicas dos smartphones estavam distribuídas e de que forma se relacionavam com o preço final.

A investigação começou pela análise da variável `price_usd`, responsável por representar o preço estimado de cada aparelho. Ao observar sua distribuição, foi possível perceber que a maior parte dos smartphones se concentrava em faixas intermediárias, enquanto valores muito baixos ou extremamente altos eram menos frequentes. Esse padrão indica um mercado predominantemente voltado para modelos de gama média, situação que reflete tendências atuais do setor, no qual aparelhos com bom equilíbrio entre desempenho e custo ocupam grande parte das vendas.

Outro ponto importante da análise foi a matriz de correlação entre as variáveis numéricas. Essa avaliação permitiu identificar que atributos como memória RAM, armazenamento interno e resolução das câmeras apresentavam relações positivas e significativas com o preço dos smartphones. Isso significa que, conforme essas especificações aumentam, o preço dos aparelhos também tende a crescer. A variável referente ao suporte ao 5G, convertida para formato numérico durante o ETL, também demonstrou influência moderada, reforçando o papel dessa tecnologia no posicionamento de mercado dos modelos analisados.

Também foi avaliado o comportamento das marcas em relação ao preço médio dos aparelhos. Essa análise mostrou que determinadas marcas se mantêm consistentemente em segmentos de preço mais elevados, enquanto outras se concentram em modelos mais acessíveis. Essa diferença ajuda a entender como as fabricantes estão distribuídas no mercado e quais delas tendem a priorizar modelos premium ou de entrada.

Gráficos gerados durante essa etapa, como histogramas, boxplots e dispersões entre variáveis, auxiliaram na identificação de outliers e na visualização de padrões relevantes. Todos esses elementos foram fundamentais para orientar a etapa de modelagem, pois permitiram identificar as variáveis com maior potencial preditivo e compreender a estrutura geral dos dados.

Ao concluir essa etapa, tornou-se possível avançar para a modelagem com uma visão clara das relações que realmente importam no comportamento dos preços. Dessa forma, a análise exploratória cumpriu seu papel de oferecer uma base sólida para a construção e avaliação dos modelos utilizados no projeto.

5. Modelagem e Avaliação dos Resultados

A etapa de modelagem teve como objetivo desenvolver modelos capazes de prever o preço dos smartphones a partir das características técnicas presentes no dataset. Para isso, foi necessário dividir os dados em dois subconjuntos, sendo um destinado ao treinamento e outro ao teste, assegurando que as avaliações realizadas posteriormente refletissem a capacidade real do modelo de generalizar para dados não vistos anteriormente.

Dois algoritmos de regressão foram escolhidos para compor essa etapa. O primeiro foi a Regressão Linear, um modelo tradicional e amplamente utilizado como referência inicial por sua simplicidade e facilidade de interpretação. Ele parte da premissa de que a relação entre as variáveis independentes e o preço pode ser representada por combinações lineares. Apesar dessa característica tornar o modelo acessível e intuitivo, ela também limita sua capacidade de capturar comportamentos mais complexos presentes no conjunto de dados.

O segundo modelo aplicado foi o Random Forest Regressor, um algoritmo composto por diversas árvores de decisão treinadas de forma independente. Cada árvore analisa subconjuntos distintos dos dados e contribui para a previsão final. Esse processo torna o modelo capaz de capturar relações

não lineares e interações sutis entre as variáveis, oferecendo maior precisão em cenários com múltiplos fatores influenciando o resultado.

Para avaliar o desempenho dos modelos, foram utilizadas as métricas RMSE e R^2 . O RMSE representa o erro médio das previsões em relação aos valores reais, enquanto o R^2 indica a proporção de variância explicada pelo modelo. A análise dessas métricas mostrou que a Regressão Linear apresentou desempenho razoável, porém limitado pela complexidade do problema. Em contraste, o Random Forest obteve valores mais elevados de R^2 e erros menores, demonstrando melhor capacidade de prever o preço dos smartphones.

Com base nesses resultados, o Random Forest foi selecionado como o modelo final do estudo. Após a seleção, ele foi serializado e salvo, permitindo sua reutilização futura em análises, aplicações práticas ou demonstrações. Essa escolha garantiu que o projeto utilizasse um modelo robusto, eficiente e compatível com as exigências da tarefa proposta.

6. Conclusão

A análise exploratória permitiu identificar que variáveis como memória RAM, armazenamento interno, câmeras e suporte ao 5G exercem influência direta sobre o preço dos aparelhos. Além disso, diferenças entre marcas destacaram o papel do posicionamento comercial das empresas na formação do valor final dos dispositivos. Essas observações ajudaram a definir expectativas mais claras para a etapa de modelagem.

Na fase de modelagem, a comparação entre Regressão Linear e Random Forest mostrou que o segundo modelo se adequa melhor às particularidades do dataset. A capacidade do Random Forest de capturar relações não lineares e interações complexas resultou em desempenho superior nas previsões, tornando-o o modelo selecionado para representar o comportamento dos dados.

O projeto conclui que características técnicas têm impacto direto e significativo sobre o preço dos smartphones. Além disso, demonstra que técnicas de aprendizado de máquina podem ser utilizadas para compreender padrões de mercado e construir modelos capazes de apoiar tomadas de decisão. Como trabalhos futuros, recomenda-se explorar métodos de otimização de hiperparâmetros, aplicar modelos mais avançados, como XGBoost, e utilizar métodos de interpretabilidade para compreender de forma mais detalhada a contribuição individual de cada variável no preço final.

7. Referências

GERON, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. 3. ed. Sebastopol: O'Reilly Media, 2022.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data Mining: Concepts and Techniques*. 3. ed. Burlington: Morgan Kaufmann, 2011.

KIMBALL, Ralph; ROSS, Margy. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3. ed. New Jersey: John Wiley & Sons, 2013.

MATPLOTLIB DEVELOPMENT TEAM. *Matplotlib Documentation*. Disponível em: <https://matplotlib.org>. Acesso em: 02 nov. 2025.

MURPHY, Kevin P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012.

PANDAS DEVELOPMENT TEAM. *Pandas Documentation*. Disponível em: <https://pandas.pydata.org>. Acesso em: 13 nov. 2025.

SCIKIT-LEARN DEVELOPERS. *Scikit-Learn Documentation*. Disponível em: <https://scikit-learn.org/stable/>. Acesso em: 02 nov. 2025.

SHAHZADI, Ayesha. *World Smartphone Market 2025 – Global Mobile Prices Extended*. Kaggle, 2024. Disponível em: <https://www.kaggle.com/datasets/shahzadi786/world-smartphone-market-2025>. Acesso em: 10 nov. 2025.

VAN ROSSUM, Guido; DRAKE, Fred L. *Python Tutorial*. Python Software Foundation, 2011. Disponível em: <https://docs.python.org/3/>. Acesso em: 02 nov. 2025.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A.; PAL, Christopher J. *Data Mining: Practical Machine Learning Tools and Techniques*. 4. ed. Cambridge: Morgan Kaufmann, 2016.