Relatório Experimental: Análise de Dados do *Dataset* Diamantes

Isabella Rodrigues de Oliveira[1]

[1]Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie (UPM)
São Paulo – SP – Brasil

10357696@mackenzista.com.br

Resumo. *Este trabalho aborda a tarefa de precificação de diamantes através do desenvolvimento de modelos de regressão com Aprendizado de Máquina (AM). Utilizando o dataset público Diamonds, foi realizado um pré-processamento de dados que incluiu a remoção de outliers e a codificação de variáveis categóricas. Diversos algoritmos, como Regressão Linear Múltipla, SVM e Árvores de Decisão, serão treinados e avaliados com base na métrica RMSE. O objetivo final é construir um modelo preditivo robusto para fornecer uma avaliação consistente do valor de diamantes. Projeto publicamente disponível na plataforma GitHub*[1].

1. Introdução

A precificação de diamantes é uma tarefa de alta complexidade, tradicionalmente dependente da avaliação de especialistas sobre um conjunto de atributos, como os ”4 Cs”: quilate (*carat*), lapidação (*cut*), cor (*color*) e pureza (*clarity*). A relação não-linear entre essas características torna o processo suscetível a inconsistências. Neste cenário, o Aprendizado de Máquina (AM) surge como uma ferramenta para superar essas limitações, utilizando algoritmos para analisar dados em larga escala e construir modelos de predição que trazem maior objetividade e precisão para o mercado.

O presente trabalho se dedica a aplicar essa abordagem, tendo como objetivo principal desenvolver e avaliar modelos de regressão para prever o preço de diamantes. Utilizando um *dataset* público, será conduzido um processo de análise e tratamento de dados que abrange desde a codificação de variáveis categóricas e normalização até a remoção de *outliers* que poderiam distorcer os resultados. Por meio da aplicação e comparação de diferentes algoritmos de AM, busca-se construir um modelo preditivo robusto que sirva como uma ferramenta de apoio à precificação nesta indústria.

2. Fundamentação Teórica

Para suavizar o processo de ajuste dos modelos de predição, algumas técnicas de pré-processamento são comumente adotadas. A remoção de dados incongruentes é importante para auxiliar o processo de interpretação dos algoritmos de AM; isso inclui dados corrompidos e *outliers*. A análise de correlações identifica quais atributos estão fortemente correlacionados com o atributo objetivo (preço) e pode justificar a remoção ou tratamento adicional de uma coluna inteira de atributos. O processo de codificação de atributos categóricos consiste em transformar dados textuais e, possivelmente, sem relações hierárquicas em dados numéricos, sendo vital para o seu uso em algoritmos de AM.

[1]https://github.com/bellaofficial23/Projeto_AI/tree/main

3. Descrição do Problema

A avaliação e precificação de pedras de diamante possuem desafios significativos para geminologistas devido à variação das características das pedras. O presente trabalho visa, por meio do uso de AM, facilitar a predição dos preços de pedras de diamante baseadas em características como: quilate, proporções, clareza, dentre outros.

4. Aspectos Éticos

Por consequência da adoção de modelos de linguagem em larga escala e uso de dados pessoais no meio digital, leis foram criadas para proteger indivíduos e suas informações. Como exemplo, temos a Lei Geral de Proteção de Dados (LGPD) no Brasil e a *General Data Protection Regulation* (GDPR) na União Europeia. Porém, o presente projeto não trabalha com dados pessoais de indivíduos e, portanto, não demanda preocupações relacionadas às regulações citadas anteriormente ou tratamentos de anonimização dos dados.

No que tange ao uso do sistema proposto, é importante notar que os artefatos a serem gerados possuirão limitações inerentes. Na possível implementação do modelo em um sistema de predição, é importante que sua documentação seja transparente quanto aos algoritmos e *dataset* utilizados para a compreensão do usuário, de forma que a interpretação final seja feita de forma consciente.

5. Dataset

O conjunto de dados escolhido para o projeto é o conjunto *Diamonds*, disponível na plataforma Kaggle [Agrawal 2025]. Esse *dataset* é composto de 53.940 instâncias e 10 variáveis, são elas: *price, carat, cut, color, clarity, length, width, depth* e *table*.

A variável *price* representa o preço em dólares americanos ($326–$18,823), *carat* representa o peso do diamante (0.2–5.01), *cut* representa a qualidade do diamante *(Fair, Good, Very Good, Premium* e *Ideal)*, *color* representa a cor do diamante (de J a D em ordem de qualidade ascendente), *clarity* representa a clareza do diamante (I1, SI2, SI1, VS2, VS1, VVS2, VVS1 e IF), x representa a altura em mm (0–10.74), y representa a largura em mm (0–58.9), z representa a profundidade em mm (0–31.8), *depth* representa a profundidade total em porcentagem de profundidade sendo a fórmula: $2z/(x+y)$ (43–79), e *table* representa a largura da mesa no topo do diamante em relação ao ponto mais largo (43–95).

5.1. Análise e Pré-Processamento

Primeiramente, foram utilizadas as seguintes bibliotecas: Pandas [pandas development team 2020] para trabalhar com dados tabulares, NumPy [Harris et al. 2020] para manipular dados em formato de arranjos e realizar operações numéricas eficientemente, Matplotlib [Hunter 2007] para criação de gráficos e scikit-learn [Buitinck et al. 2013] que oferece ferramentas para análise de dados e modelagem preditiva.

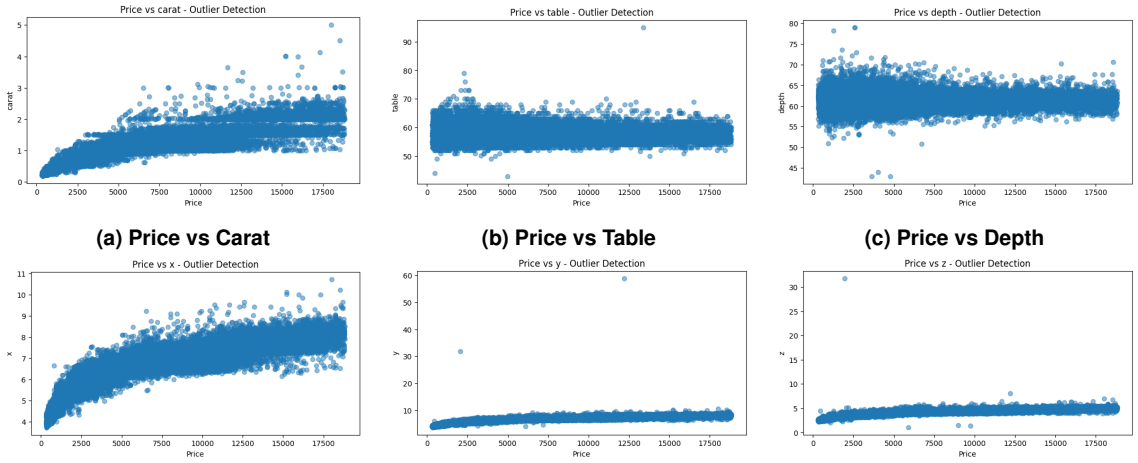
Iniciando o pré-processamento de dados, foi identificado que a primeira coluna da tabela continha dados referentes ao número das instâncias, o que se mostrou irrelevante para a análise e, portanto, a coluna foi removida. Após isso, por meio do método *describe* foi descoberta a existência de valores mínimos iguais a 0 nas colunas x, y e z, visíveis na

	carat	*depth*	*table*	*price*	x	y	z
Média	0.798	61.749	57.457	3932.799	5.731	5.735	3.539
DP	0.474	1.433	2.234	3989.439	1.122	1.142	0.705
Mínimo	0.2	43.0	43.0	326.0	0.0	0.0	0.0
Máximo	5.01	79.0	95.0	18823.0	10.740	58.90	31.80

Tabela 1. Informações resultantes do método *describe* do *dataframe* Pandas.

Tabela 5.1. O que não faz sentido, visto que esses atributos dizem respeito às dimensões do diamante, portanto, tais instâncias também foram removidas.

Posteriormente, foi realizada a codificação de dados categóricos (sendo as variáveis *cut*, *clarity* e *color*) em dados numéricos a fim de estarem aptos à interpretação por algoritmos de AM. Após isso, análises gráficas determinaram a existência de *outliers* a serem removidos (Figura 1). O número de instâncias do dataset após todas as remoções foi de 53910.



(a) Price vs Carat **(b) Price vs Table** **(c) Price vs Depth**

(d) Price vs X **(e) Price vs Y** **(f) Price vs Z**

Figura 1. Análise gráfica dos atributos numéricos para identificação de *outliers*.

Com a análise de correlações, chegamos à conclusão de que a variável mais influente no preço é a *carat*. Logo, o dataset foi separado em "baldes"de acordo com a variável *carat* a fim de termos uma separação estratificada dos conjuntos de treino e teste.

6. Metodologia

Para desenvolver e avaliar os diversos métodos de predição, será utilizada a fórmula de Erro Quadrático Médio (RMSE), que informa a quantia de erros presentes no sistema desenvolvido e é descrita pela Equação $\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$ [Géron 2017]. Dentre os algoritmos a serem avaliados, estão: Support Vector Machine (SVM), Árvore de Decisão, Redes Neurais, Regressão Linear Múltipla e K-Vizinhos Mais Próximos (KNN).

Espera-se que, por meio do "*input*" do usuário, o modelo retorne uma predição que sirva de base para a precificação de um diamante. Espera-se também que o modelo demonstre bom ajuste aos dados (sem *overfitting* ou *underfitting*) e que, dadas as métricas

de erro e performance (RMSE e coeficiente de determinação), seu desempenho seja satisfatório.

Referências

Agrawal, S. (2025). Kaggle: diamonds dataset.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn & TensorFlow CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS*. O'Reilly, Boston - Farnham - Sebastopol - Tokyo - Beijing.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

pandas development team, T. (2020). pandas-dev/pandas: Pandas.