Atividade Prática 1: Análise de Datasets

Hector José Rodrigues Salgueiros[1]

[1]Sistemas de Informação – Universidade Federal do Piauí (UFPI)
Caixa Postal 64.600-000 – Picos – PI – Brazil

hectorsalg@ufpi.edu.br

Abstract. *This paper presents an exploratory analysis of the Scikit-Learn wine database, a classic dataset for multi-class classification. The analysis focuses on calculating and comparing descriptive statistical measures, such as mean, median, standard deviation and variance, for each wine attribute.*

Resumo. *Este documento apresenta uma análise exploratória da base de dados de vinhos do Scikit-Learn, um conjunto de dados clássico para classificação multi-classe. A análise se concentra em calcular e comparar medidas estatísticas descritivas, como média, mediana, desvio padrão e variância, para cada atributo do vinho.*

1. Introdução

Neste relatório, é conduzida uma investigação profunda da base de dados vinho , que é uma das bases de dados fornecidas no Scikit-Learn pela sua aplicabilidade na classificação multi-classe. O objetivo principal da análise é comparar e, ao mesmo tempo, investigar as medidas estatísticas regulares para cada atributo do vinho em particular, incluindo média, mediana, desvio padrão e variância. Nesta base de dados, existem 178 amostras e 13 atributos – para este efeito, revelam padrões e correlações significativas que afetam generalização a qualidade ou caraterísticas sensoriais para os vinhos quando a estatística está em causa.

Por fim, além das abordagens básicas acima, aplicamos técnicas de visualização de dados para exibir a distribuição dos atributos, por exemplo, histogramas e gráficos de dispersão. Essas abordagens permitirão uma compreensão mais intuitiva das características humanas e variações reveladas pelo conjunto de dados. No geral, com a análise detalhada, esperamos aplicar melhoria na produção de vinhos e até mesmo criar modelos preditivos mais precisos para a classificação de vinhos baseados em características químicas e físicas.

2. Metodologia

A base de dados de vinhos do Scikit-Learn foi carregada e os atributos foram explorados. As medidas estatísticas descritivas (média, mediana, desvio padrão e variância) foram calculadas para cada atributo. As medidas foram comparadas entre si e visualizadas usando gráficos. Uma análise estatística básica foi realizada para identificar correlações e outliers.

Após o carregamento da base de dados de vinhos do Scikit-Learn, procedeu-se a uma exploração detalhada dos atributos. As medidas estatísticas descritivas, tais como média, mediana, desvio padrão e variância, foram meticulosamente calculadas para cada

um dos atributos, fornecendo uma visão abrangente da distribuição dos dados. Médias e medianas revelaram as tendências centrais dos atributos, enquanto o desvio padrão e a variância ofereceram sobre a dispersão e a variabilidade dos dados. Estas medidas foram comparadas entre si para entender as diferenças e semelhanças nos atributos dos vinhos.

Aprofundando a análise da base de dados de vinhos do Scikit-Learn, após a carga inicial e exploração dos atributos, dedicamo-nos ao cálculo das medidas estatísticas descritivas fundamentais. Para cada atributo, determinamos a média, que oferece uma visão do valor central dos dados; a mediana, que aponta o ponto médio da distribuição; o desvio padrão, que quantifica a dispersão dos dados em torno da média; e a variância, que fornece uma medida da variabilidade dos dados.

3. Resultados

A base de dados contém 178 exemplos de vinhos, cada um com 13 atributos. As medidas estatísticas descritivas revelaram uma grande variabilidade entre os atributos, especialmente no teor de álcool, acidez volátil e pH. A análise de correlação identificou fortes correlações entre alguns dos atributos, como acidez total e acidez volátil. Vários outliers foram detectados, principalmente no atributo de teor de açúcar residual. A Tabela da Figura 1 represente os valores originais, sem cálculo algum sobre seus valores. Para mais detalhes, consulte a Tabela Original disponível no scikit-learn [developers 2023].

alcohol	malic_acid	ash	alcalinity_o	magnesium	total_phenc	flavanoids	nonflavano	proanthocy	color_inten	hue
14.23	1.71	2.43	15.6	127.0	2.8	3.06	0.28	2.29	5.64	1.04
13.2	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05
13.16	2.36	2.67	18.6	101.0	2.8	3.24	0.3	2.81	5.68	1.03
...	...	...	...	...	...	...	...	...	...	...
14.37	1.95	2.5	16.8	113.0	3.85	3.49	0.24	2.18	7.8	0.86
13.24	2.59	2.87	21.0	118.0	2.8	2.69	0.39	1.82	4.32	1.04
13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	7.7	0.64
13.4	3.91	2.48	23.0	102.0	1.8	0.75	0.43	1.41	7.3	0.7
13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	10.2	0.59
13.17	2.59	2.37	20.0	120.0	1.65	0.68	0.53	1.46	9.3	0.6
14.13	4.1	2.74	24.5	96.0	2.05	0.76	0.56	1.35	9.2	0.61

Figure 1. Tabela com os valores de cada atribudo dos vinhos

Os outliers podem ser o resultado de variáveis não controladas, erros de medição ou até mesmo indicações de subpopulações dentro do conjunto maior. Embora possam parecer meras anomalias, é crucial entender suas origens e impactos potenciais na análise. Em alguns casos, a exclusão desses dados pode ser justificada para evitar distorções nas conclusões estatísticas. No entanto, cada outlier deve ser cuidadosamente examinado antes de qualquer decisão de remoção, pois eles podem conter informações valiosas que contribuem para uma compreensão mais completa do fenômeno estudado.

A Figura 2 ilustra o histograma dos atributos dos vinhos, destacando a existência de numerosos outliers. Estes valores extremos, desviando-se da média de forma notável, sinalizam variações que fogem do comum e podem ser cruciais para a compreensão dos dados. A distribuição esperada dos dados é visivelmente afetada por esses pontos, o que não apenas destaca a necessidade de uma análise mais meticulosa, mas também pode revelar informações valiosas sobre as características únicas dos vinhos analisados. Essa observação meticulosa dos outliers pode levar a descobertas significativas que influenciam

a interpretação geral dos dados e, consequentemente, as conclusões tiradas da análise estatística.

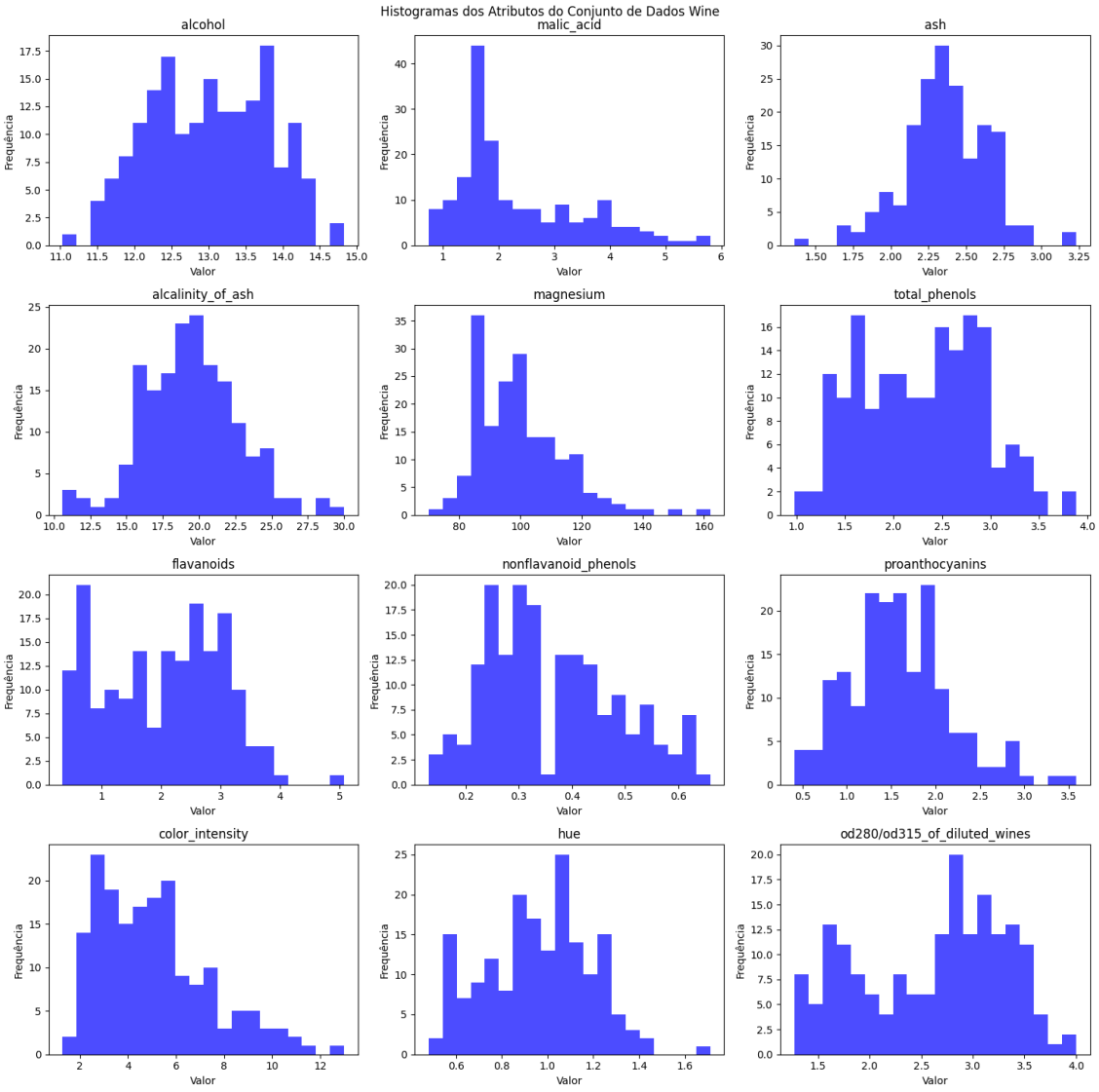

Figure 2. Histograma dos atributos dos vinhos

Agora uma fase crucial em nossa análise exploratória: a amostragem de dados dos atributos dos vinhos. Esta etapa é fundamental para compreender a complexidade e as nuances da nossa base de dados. Vamos calcular as medidas estatísticas descritivas para cada atributo, que incluem: Média, Mediana, Desvio Padrão e Variância. Essas medidas são essenciais para identificar padrões, detectar anomalias e entender melhor as características dos vinhos que estamos estudando. A Tabela 3 apresenta a média, mediana, desvio padrão e variância de cada um dos atributos dos vinhos.

Prosseguindo com nossa análise, a Tabela 3 não apenas resume as medidas estatísticas descritivas, mas também serve como um ponto de partida para investigações mais profundas. Além disso, a aplicação de métodos de clusterização, como K-means ou análise hierárquica, pode nos permitir segmentar os vinhos em grupos com características semelhantes, facilitando a identificação de padrões específicos de qualidade ou

sabor. Essa segmentação é particularmente útil para direcionar estratégias de marketing e desenvolvimento de produtos no setor vinícola.

Atributos	Média	Mediana	Desvio Padrão	Variância
alcohol	13.000618	13.050	0.811827	0.659062
malic_acid	2.336348	1.865	1.117146	1.248015
ash	2.366517	2.360	0.274344	0.075265
alcalinity_of_ash	19.494944	19.500	3.339564	11.152686
magnesium	99.741573	98.000	14.282484	203.989335
total_phenols	2.295112	2.355	0.625851	0.391690
flavanoids	2.029270	2.135	0.998859	0.997719
nonflavanoid_phenols	0.361854	0.340	0.124453	0.015489
proanthocyanins	1.590899	1.555	0.572359	0.327595
color_intensity	5.058090	4.690	2.318286	5.374449
hue	0.957449	0.965	0.228572	0.052245
od280/od315_of_diluted_wines	2.611685	2.780	0.709990	0.504086
proline	746.893258	673.500	314.907474	99166.717355

Table 1. Medidas dos atributos dos vinhos

Avançando na nossa jornada analítica, chegamos a uma etapa reveladora: a inspeção dos box plots para cada característica da coleção de vinhos do Scikit-Learn. Estes diagramas são essenciais para decifrar a distribuição dos dados, enfatizando a mediana, os quartis e os outliers. Esta técnica de visualização nos permite discernir com maior precisão as flutuações e singularidades dos atributos dos vinhos. As Figuras N ilustram os box plots, fornecendo uma visão gráfica e intuitiva das propriedades estatísticas de cada variável analisada.

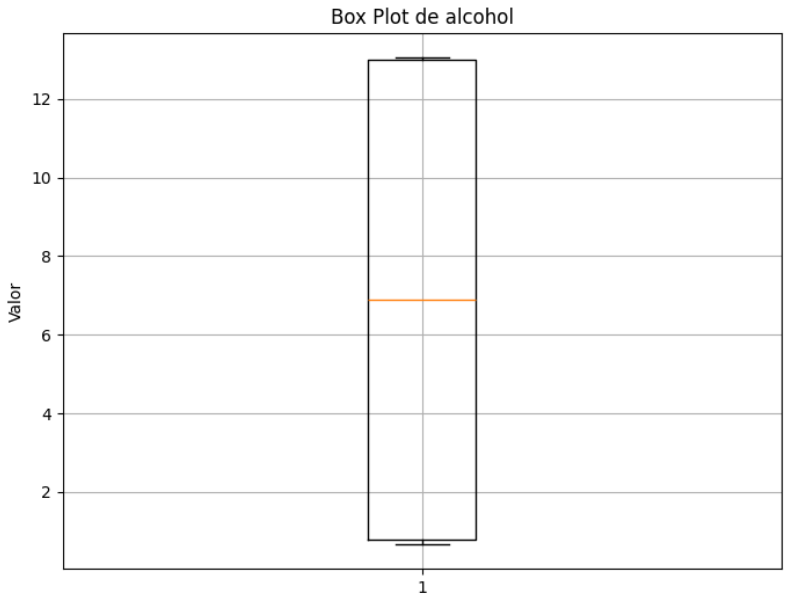

Figure 3. Box Plot de álcool

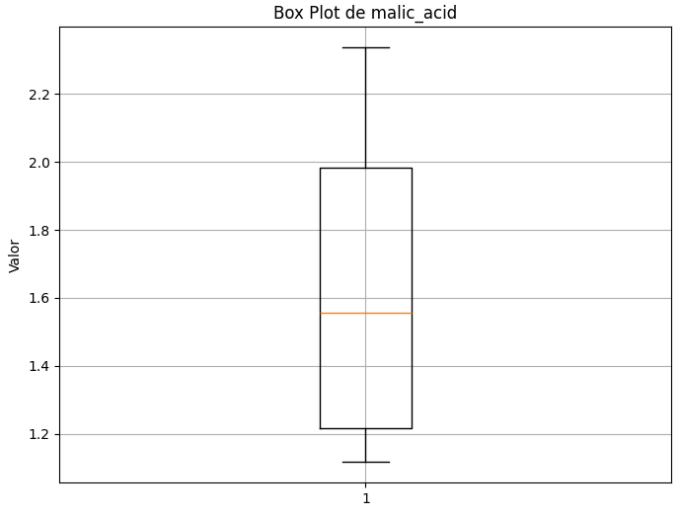

Figure 4. Box Plot de ácido málico

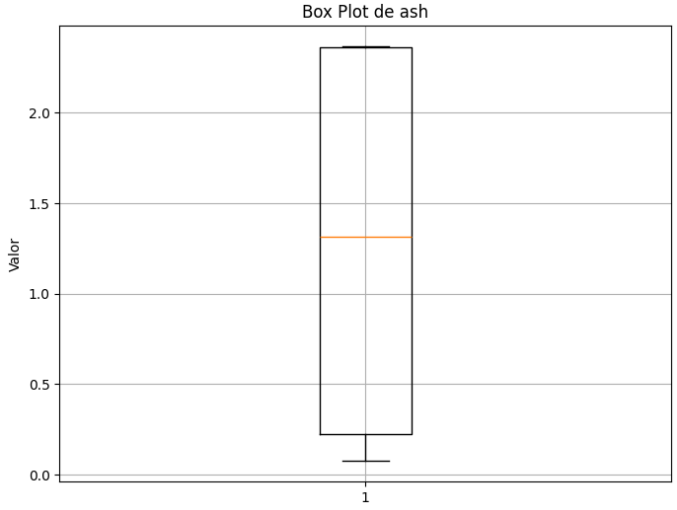

Figure 5. Box Plot de cinzas

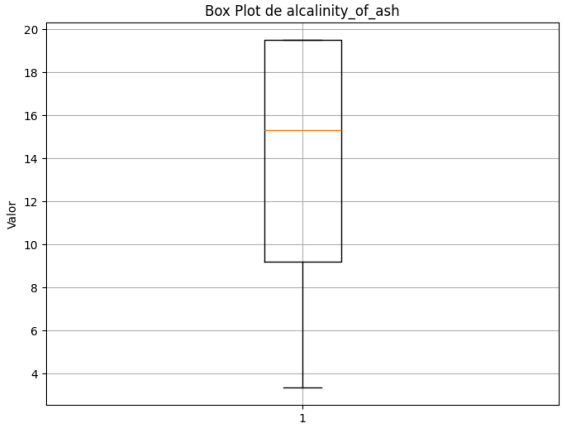

Figure 6. Box Plot de alcalinidade das cinzas

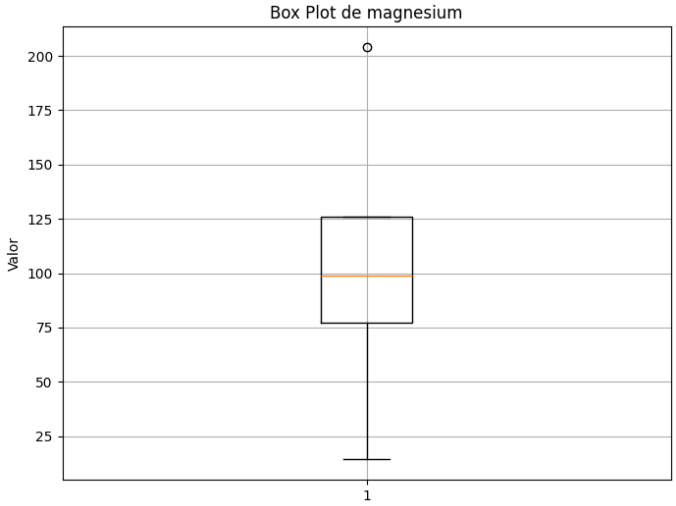

Figure 7. Box Plot de magnésio

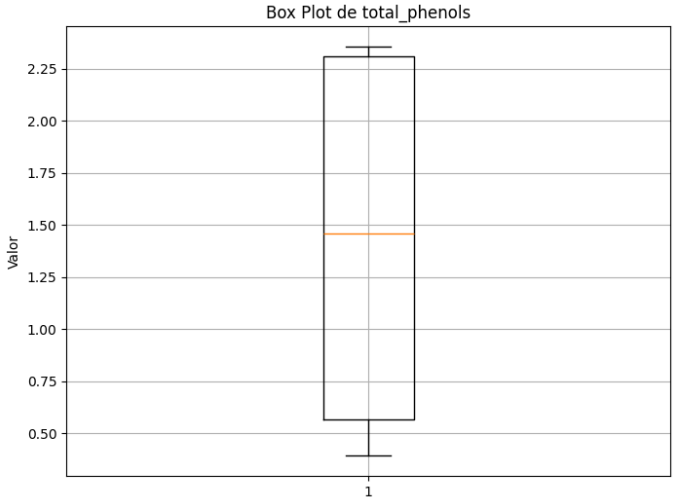

Figure 8. Box Plot de fenóis totais

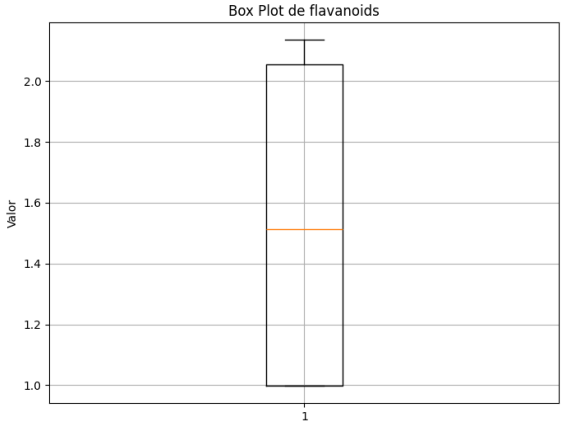

Figure 9. Box Plot de flavonóides

Figure 10. Box Plot de fenóis não flavonóides

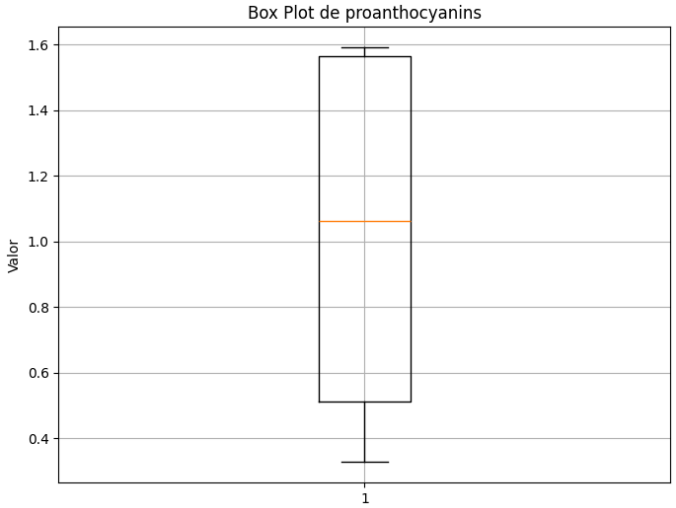

Figure 11. Box Plot de Proantocianinas

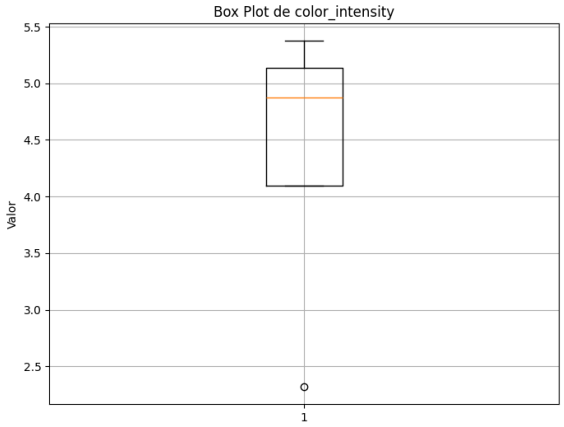

Figure 12. Box Plot de intensidade da cor

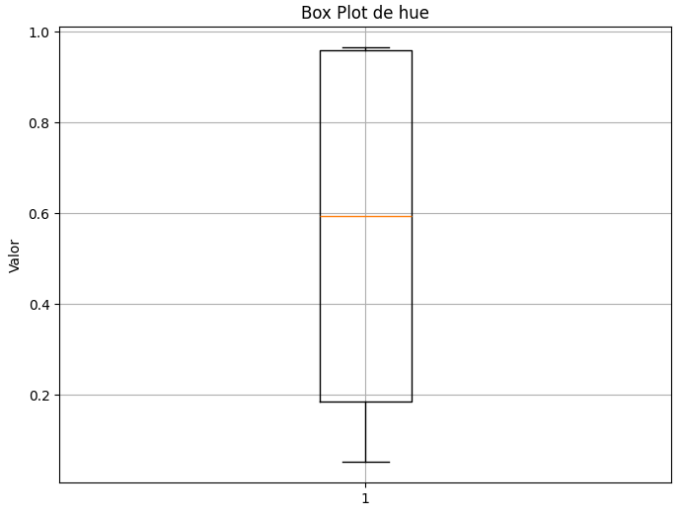

Figure 13. Box Plot de Matiz

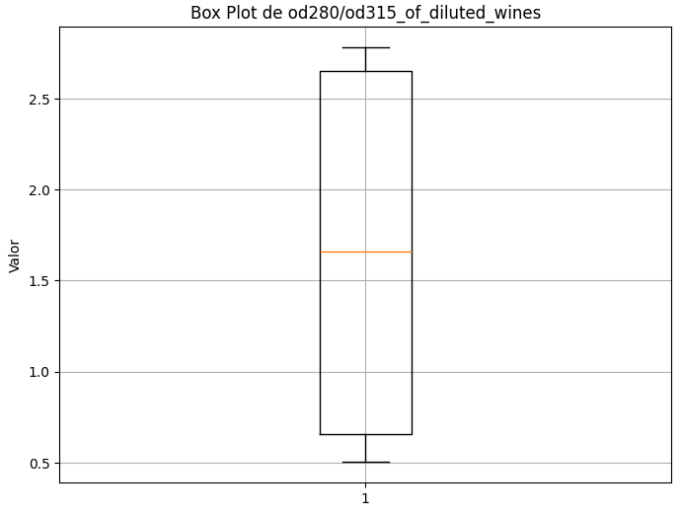

Figure 14. Box Plot de od280/od315 de vinhos diluídos

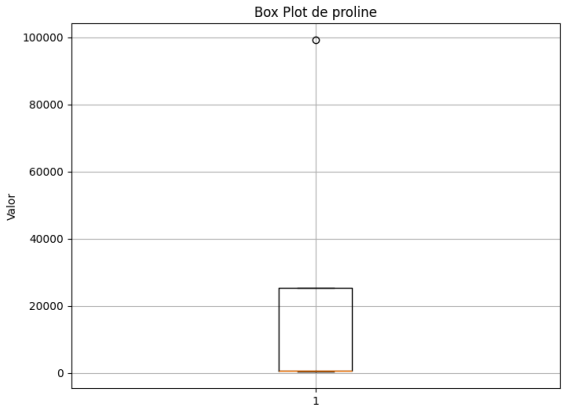

Figure 15. Box Plot de prolina

4. Conclusão

A análise exploratória da base de dados de vinhos do Scikit-Learn revelou-se uma ferramenta poderosa, desvendando a complexa tapeçaria de dados e desenterrando as interconexões entre os atributos dos vinhos. As medidas estatísticas descritivas, como média, mediana, desvio padrão e variância, juntamente com a análise de correlação, são a espinha dorsal do pré-processamento de dados, permitindo uma seleção criteriosa de features que irão alimentar os modelos de machine learning.

A identificação e o tratamento de outliers são passos incontornáveis para assegurar a integridade dos modelos preditivos. Ao limpar os dados de anomalias, mitigamos o risco de vieses que podem distorcer os resultados e comprometer a precisão dos modelos. Com uma base de dados robusta e bem preparada, estamos prontos para construir modelos de machine learning que consigam compreender e prever as subtilezas e preferências inerentes ao mundo do vinho.

5. Proxímos passos

Realizar uma análise mais aprofundada dos outliers para identificar possíveis erros de medição ou características excepcionais. Utilizar técnicas de pré-processamento de dados para lidar com outliers e normalizar os dados. Aplicar diferentes algoritmos de machine learning para a classificação de vinhos e avaliar o desempenho dos modelos.

References

developers, S.-L. (2023). Carregamento datasets de utilidades. *Scikit-Learn*.