Relatório I - Tópicos Especiais em Computação

Vandirleya Barbosa da Costa

[1]Universidade Federal de Piauí (UFPI)

vandirleya.barbosa,israel.araujo@ufpi.edu.br

Resumo. *Este trabalho apresenta uma análise realizada sobre o conjunto de dados Wine, amplamente utilizado no campo do aprendizado de máquina e originário do UCI Machine Learning Repository. O conjunto de dados sobre vinhos é tabular e foi desenvolvido visando utilizar análise química para determinar a origem de vinhos. Composto por 178 instâncias, possui 13 características que representam diferentes aspectos químicos dos vinhos. Os dados são resultados de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivares diferentes. Essa análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. O objetivo da análise realizada foi conduzir uma análise descritiva dessa base e identificar padrões ou tendências que possam indicar características importantes relacionadas à qualidade do vinho. Portanto, este estudo propôs análises estatísticas, bem como um resumo exploratório dos resultados obtidos, visando compreender melhor as relações entre as características químicas dos vinhos e sua origem ou qualidade.*

Abstract. *This paper presents an analysis conducted on the Wine dataset, widely used in the field of machine learning and originating from the UCI Machine Learning Repository. The dataset on wines is tabular and was developed with the aim of using chemical analysis to determine the origin of wines. Comprising 178 instances, it has 13 features representing different chemical aspects of wines. The data is the result of chemical analysis of wines grown in the same region of Italy, but derived from three different cultivars. This analysis determined the quantities of 13 constituents found in each of the three types of wines. The objective of the analysis was to conduct a descriptive analysis of this dataset and identify patterns or trends that may indicate important characteristics related to wine quality. Therefore, this study proposed statistical analyses, as well as an exploratory summary of the results obtained, aiming to better understand the relationships between the chemical characteristics of wines and their origin or quality.*

1. Introdução

O uso de aprendizado de máquina na análise de vinhos tem revolucionado a forma como entendemos e apreciamos essa bebida milenar. Tradicionalmente, a degustação de vinhos é uma habilidade complexa que requer anos de treinamento e experiência para dominar. No entanto, com o avanço da tecnologia, especialmente no campo da inteligência artificial e aprendizado de máquina, agora é possível empregar algoritmos sofisticados

para analisar características sensoriais, prever a qualidade e até mesmo identificar variedades específicas de vinho com precisão impressionante. A análise de dados desempenha um papel fundamental em diversos campos, incluindo o aprendizado de máquina.

No contexto do vinho, a análise química oferece percepções importantes sobre sua origem, composição e qualidade. O conjunto de dados Wine, referenciado como [Aeberhard and Forina 1991], é uma ferramenta amplamente utilizada para explorar esses aspectos. Neste relatório, é conduzida uma análise descritiva das principais variáveis em um conjunto de dados de vinhos, o que é crucial para compreender sua composição e potencial qualidade. Ao utilizar técnicas estatísticas e visualizações, como medidas de tendência central e dispersão, histogramas, box plots e gráficos de dispersão, podemos obter visões valiosas sobre as características dos vinhos e suas relações.

2. Metodologia

Esta seção descreve a metodologia adotada para a elaboração deste relatório. Serão detalhados o uso do conjunto de dados e os passos seguidos para a análise e obtenção dos resultados.Isso inclui a seleção das variáveis de interesse, a aplicação de técnicas estatísticas específicas e a seleção dos gráficos para interpretação dos padrões identificados.

2.1. Códigos

A princípio uma análise foi realizado no no conjunto de dados do Vinho, utilizando a biblioteca Scikit-learn, Pandas para manipulação de dados, Matplotlib e Seaborn para visualização. O objetivo é explorar estatisticamente as características químicas dos vinhos e entender como elas se relacionam com as diferentes classificações de vinhos presentes no conjunto de dados. O primeiro passo foi carregar o conjunto de dados utilizando a função load_wine() do Scikit-learn e transformá-lo em um DataFrame do Pandas para facilitar a manipulação. Além disso, a coluna de rótulos (target), que indica a classificação do vinho, foi adicionada ao DataFrame para permitir análises futuras. Em seguida, foi realizada uma análise descritiva das variáveis, incluindo a geração de estatísticas como média, desvio padrão, mínimo, máximo e mediana.

A mediana foi calculada separadamente e adicionada ao resumo estatístico para oferecer uma visão completa da distribuição dos dados.Para uma melhor compreensão da distribuição das variáveis, foram gerados histogramas para cada atributo. Os histogramas ajudam a visualizar a frequência de diferentes valores de características, indicando distribuições, tendências e possíveis outliers. Adicionalmente, box plots foram criados para cada característica em relação às diferentes classificações de vinhos (rótulos). Isso permite identificar diferenças nas distribuições de características entre os tipos de vinho e destacar possíveis valores atípicos. Por último, um gráfico de dispersão foi utilizado para examinar as relações entre todas as variáveis, com a classificação do vinho como hue. Essa visualização é essencial para identificar correlações potenciais e padrões distintos entre os diferentes tipos de vinho. O código utilizado se encontra em anexo junto com este documento.

3. Resultados

Esta seção apresenta uma discussão acerca dos resultados obtidos. A Tabela 1 apresenta estatísticas descritivas dos diferentes tipos de vinho presentes no conjunto de dados. Cada

linha representa uma característica química específica do vinho, enquanto as colunas fornecem a média, mediana, desvio padrão e variância dessas características para cada tipo de vinho. As estatísticas médias indicam o valor médio de cada característica para os diferentes tipos de vinho. Por exemplo, a média do Alcohol é de aproximadamente 13%, enquanto a média da Malic Acid é de cerca de 2.34. A mediana, por sua vez, representa o valor central de cada característica para os diferentes tipos de vinho. Ela é menos sensível a valores extremos do que a média e fornece uma medida mais robusta da tendência central dos dados. O desvio padrão é uma medida de dispersão que indica o quanto os valores individuais estão distantes da média. Quanto maior o desvio padrão, maior é a dispersão dos dados em torno da média. Por exemplo, a característica Magnesium possui um desvio padrão de aproximadamente 14.28, o que sugere uma variabilidade moderada nos níveis de magnésio entre os diferentes tipos de vinho. A variância é o quadrado do desvio padrão e representa a dispersão total dos dados em relação à média. Valores de variância mais altos indicam uma maior dispersão dos dados em torno da média.

Tabela 1. Estatísticas dos Tipos de Vinho.

Tipos de Vinhos	Média	Mediana	Desvio Padrão	Variância
Alcohol	13.00	13.05	0.81	0.66
Malic Acid	2.34	1.86	1.12	1.25
Ash	2.37	2.36	0.27	0.08
Alcalinity of Ash	19.49	19.50	3.34	11.15
Magnesium	99.74	98.00	14.28	204.12
Total Phenols	2.30	2.36	0.63	0.39
Flavanoids	2.03	2.14	1.00	1.00
Nonflavanoid Phenols	0.36	0.34	0.12	0.02
Proanthocyanins	1.59	1.56	0.57	0.33
Color Intensity	5.06	4.69	2.32	5.35
Hue	0.96	0.97	0.23	0.05
OD280/OD315 of Diluted Wines	2.61	2.78	0.71	0.50
Proline	746.89	673.50	314.91	99011.21

A Figura 1 apresenta histogramas que representam as características químicas dos diferentes tipos de vinho. A análise desses histogramas permite inferir sobre diversos aspectos, como o teor alcoólico, a acidez, a presença de compostos antioxidantes e a intensidade da cor. O Alcohol mostra uma concentração mais alta de álcool na maioria das amostras, isso pode indicar que os vinhos da região têm um teor alcoólico relativamente alto, o que pode ser característico de determinadas variedades de uvas ou práticas de vinificação específicas. O histograma do malic_acid sugere uma variação significativa neste componente entre as diferentes cultivares, o que pode ser influenciado por práticas de vinificação, características do solo ou condições climáticas da região de cultivo. A distribuição do Ash pode indicar diferenças no conteúdo mineral do solo de onde as uvas são cultivadas, fornecendo visões sobre a influência do terroir na composição dos vinhos. Um pico distinto no Alcalinity of Ash pode apontar para um processo de vinificação que afeta o pH do vinho, o que pode ser relevante para características como a acidez e a estabilidade do produto final. A presença de Magnesium é crucial para o desenvolvimento da videira e a sua distribuição pode refletir as práticas agrícolas da região, assim como o

potencial de influenciar características organolépticas dos vinhos. Altos níveis de Total Phenols estão frequentemente associados a vinhos de maior qualidade e maior capacidade antioxidante, podendo indicar vinhos mais estruturados e complexos.

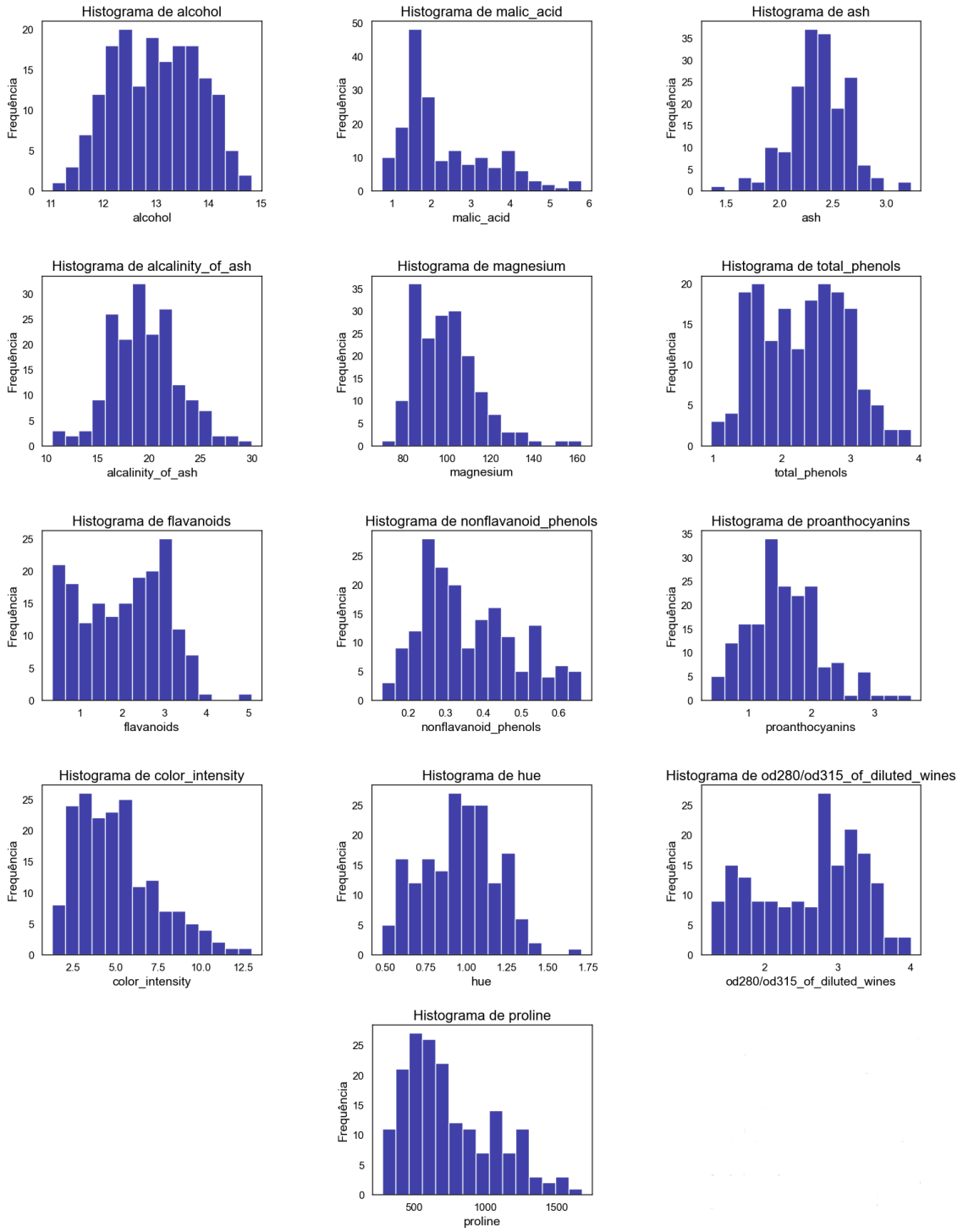

Figura 1. Histogramas dos tipos de vinho.

Flavanoids são importantes antioxidantes e sua distribuição pode indicar o potencial de envelhecimento do vinho, bem como sua capacidade de desenvolver aromas

e sabores complexos ao longo do tempo. No Nonflavanoid Phenols uma distribuição uniforme pode sugerir uma consistência na proteção contra oxidação entre as amostras, contribuindo para a preservação da qualidade e integridade do vinho. Proanthocyanins afetam a adstringência e a cor do vinho, o histograma pode mostrar como essas características variam entre as cultivares, influenciando na percepção sensorial do produto final. A variação na Color Intensity pode ser um indicativo da maturação das uvas e das técnicas de vinificação, proporcionando percepções sobre a concentração de compostos pigmentados presentes no vinho. A Hue do vinho está relacionada com a sua idade e o tipo de uva, o histograma pode revelar a gama de matizes presentes nas amostras, fornecendo informações sobre a evolução e características visuais dos vinhos ao longo do tempo. OD280/OD315 of Diluted Wines é um indicador da estabilidade e maturidade do vinho e pode mostrar a prontidão para o consumo, destacando vinhos que estão em estágio adequado para apreciação imediata ou envelhecimento adicional. Proline é Um aminoácido indicativo da qualidade do vinho e a sua distribuição pode refletir a qualidade geral das uvas e do vinho,

A Figura 2 apresenta os boxplots das variáveis analisadas. O boxplot do Alcohol indica que o Tipo de Vinho 1 tende a ter um teor alcoólico mais elevado, o que pode ser um traço distintivo dessa variedade. O Acid_malic apresenta maior dispersão no Tipo de Vinho 2 sugere uma variação significativa neste componente, possivelmente devido a diferenças nas práticas de vinificação ou no terroir. Sem diferenças significativas observáveis nos boxplots do Ash, ambos os tipos de vinho parecem ter conteúdo mineral semelhante. No Alcalinity of Ash não há picos distintos visíveis, o que torna difícil inferir qualquer processo específico de vinificação que afete o pH. No Magnesium as distribuições semelhantes em ambos os tipos não fornecem visões claras sobre as práticas agrícolas regionais. No Total Phenols o Tipo de Vinho 1 mostra uma mediana e distribuição mais altas, indicando níveis mais elevados de fenóis totais e potencialmente maior qualidade e capacidade antioxidante.

No Flavanoids assim como os fenóis totais, o Tipo de Vinho 1 também exibe uma concentração mais alta de flavonoides, sugerindo um melhor potencial de envelhecimento. Em Nonflavanoid Phenols ambos os tipos exibem distribuições bastante uniformes, no entanto, o Tipo de Vinho 2 tem uma faixa ligeiramente mais ampla, indicando mais consistência na proteção contra oxidação. No Proanthocyanins a distribuição é bastante semelhante para ambos os tipos, o que desafia a dedução de variações na cor e adstringência do vinho a partir deste gráfico. Em Color Intensity existe uma diferença notável entre os dois tipos, o Tipo de Vinho 1 geralmente exibe maior intensidade de cor, indicativa de diferentes estágios de maturação das uvas ou técnicas de vinificação empregadas. A Hue varia mais amplamente no Tipo de Vinho 2, conforme visto pelo seu maior intervalo interquartil, o que pode ser indicativo de uma diversidade de idades e tipos de uva. No OD280/OD315 of Diluted Wines o tipo de Vinho 1 é o que apresenta um indicativo de melhor estabilidade e maturidade do vinho, seguido pelo tipo de vinho 0. O Proline segue as mesmas características.

A Figura 3 apresenta um conjunto de gráficos de dispersão que ilustram as relações entre diferentes variáveis químicas em vinhos. O Alcohol apresenta uma maior concentração nos Tipos 0 e 1, sugerindo que estes podem ser vinhos com características mais robustas ou de regiões com condições climáticas específicas que favorecem um

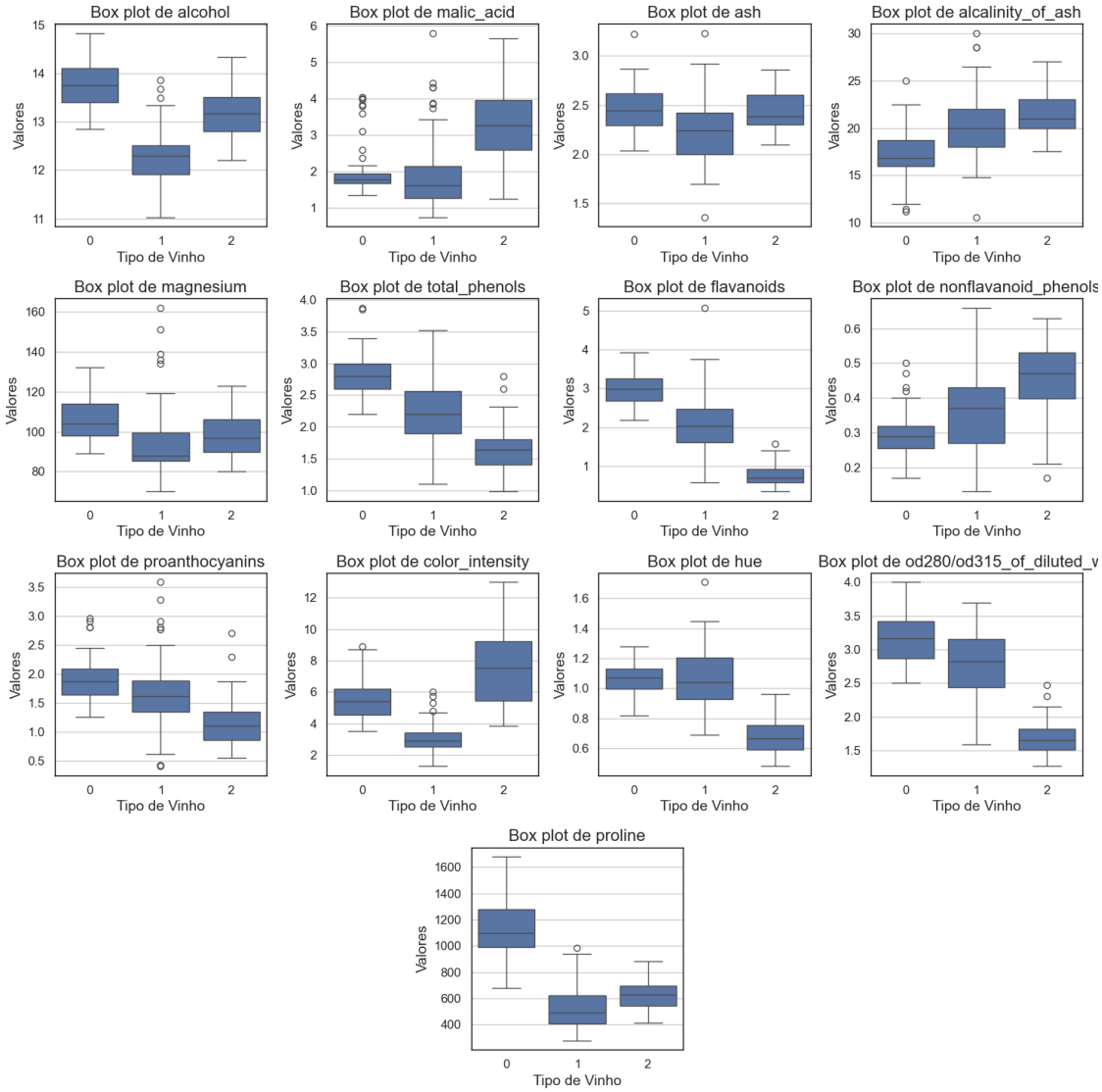

Figura 2. Box plots por vinho.

maior teor alcoólico. O Acid_malic mostra um pico de distribuição no Tipo 0, o que pode indicar uma característica distintiva dessa categoria de vinho, talvez relacionada com a escolha de uvas ou métodos de vinificação. O Ash apresenta uma concentração e densidade similares nos três tipos de vinho, refletindo uma consistência no conteúdo mineral do solo de onde as uvas são cultivadas. A mesma observação se aplica à Alcalinity of Ash, Magnesium e Total Phenols, indicando que essas propriedades são relativamente uniformes entre os diferentes tipos de vinho.

O Flavanoids apresenta uma maior dispersão no Tipo de vinho 2, o que pode ser um indicativo de variação nas práticas de vinificação ou na maturação das uvas. Por outro lado, os Nonflavanoid Phenols e Proanthocyanins apresentam uma maior dispersão no tipo 0, sugerindo uma diversidade maior nestes compostos que podem influenciar a cor e a adstringência do vinho. A Color Intensity apresenta uma maior concentração no Tipo de vinho 1, possivelmente indicando uma maior extração de pigmentos durante a vinificação ou uvas com maior intensidade de cor natural. O Hue e OD280/OD315 of

Diluted Wines mostram uma maior concentração em todos os tipos, o que pode refletir a idade e a estabilidade dos vinhos. E o Proline mostra uma concentração mais elevada nos tipos 1 e 0, associada frequentemente a vinhos de alta qualidade e regiões de clima mais frio. Essas observações podem ajudar a diferenciar os tipos de vinho e a entender as complexidades da vinificação e da composição química dos vinhos

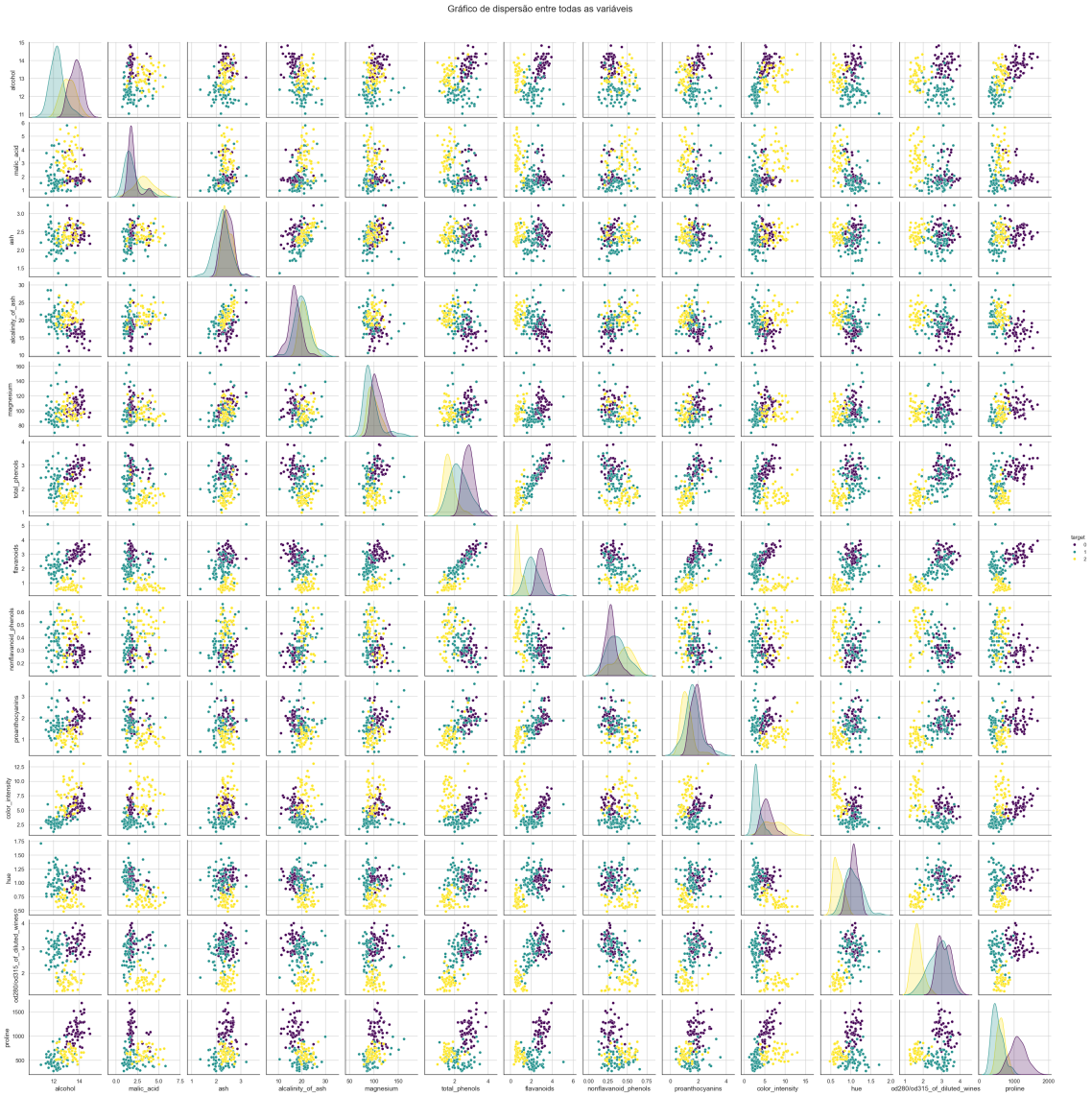

Figura 3. Gráficos de dispersão.

4. Conclusão

Este trabalho propôs a realização de uma análise de dados de um dataset de vinho para identificar padrões ou tendências que possam indicar características importantes relacionadas à qualidade do vinho. Com base na análise exploratória dos dados e nas relações identificadas entre as variáveis químicas e a classificação de qualidade do vinho, podemos concluir que existem padrões e tendências significativas que podem influenciar a qualidade e as características dos vinhos. Os resultados revelaram que certas variáveis químicas, como Alcohol, Malic_acid, presença de compostos antioxidantes e intensidade

da cor, variam entre os diferentes tipos de vinho. Por exemplo, Alcohol tende a ser mais elevado nos tipos de vinho 0 e 1, sugerindo características mais robustas ou condições climáticas específicas favoráveis a um maior teor alcoólico. Além disso, a presença de compostos como Flavanoids e Total Phenols pode indicar maior potencial de envelhecimento e qualidade do vinho. A análise também revelou que algumas variáveis, como a Color Intensity e a concentração de Proline, estão associadas a determinados tipos de vinho, o que pode fornecer informações sobre a origem das uvas ou as práticas de vinificação empregadas. Esses padrões e correlações identificados podem ser úteis para produtores de vinho e enólogos na seleção de uvas, no desenvolvimento de técnicas de vinificação e no aprimoramento da qualidade dos vinhos.

Referências

Aeberhard, S. and Forina, M. (1991). Wine. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC7J.