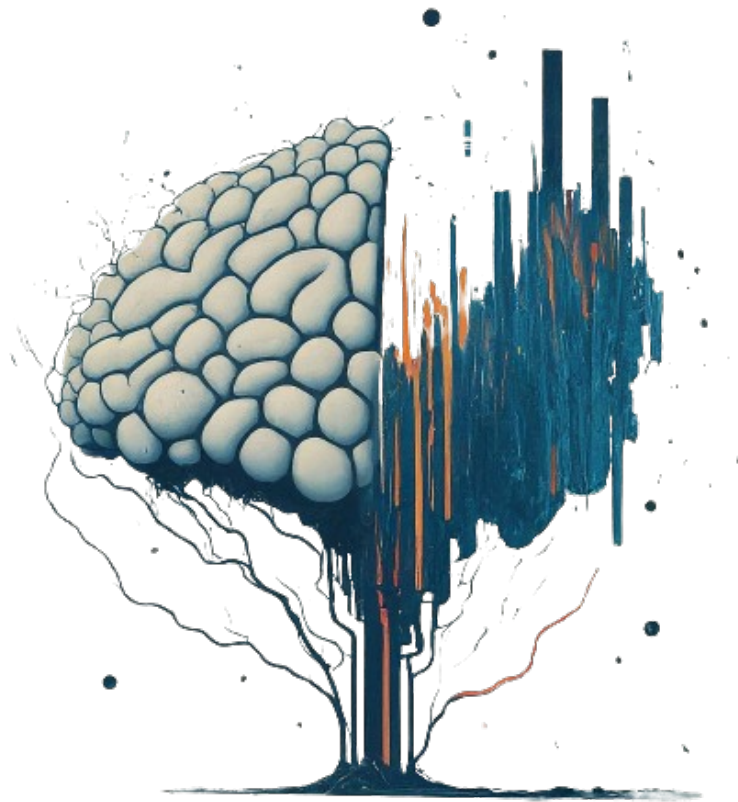


Análise de Dados de Amostras de Beterraba

Análise Inteligente de Dados

Instituto Politécnico de Coimbra
Instituto Superior de Engenharia de Coimbra
Licenciatura em Engenharia Biomédica



Andreia Fernandes-20202118612
Rita Quaresma-2022143948

Índice

Resumo	1
1 Introdução	2
2 Metodologia	2
2.1 Etapas do CRISP-DM	2
2.2 Compreensão do problema	4
2.2.1 Estado da Arte na Análise de Dados de Amostras de Be- terraba	4
2.3 Compreensão dos dados	5
2.3.1 Variáveis de Análise	5
2.3.2 Objetivos do Trabalho:	8
2.3.3 Código usado no Python nesta etapa:	8
2.4 Preparação dos dados	11
2.4.1 Boxplot	12
2.4.2 Histograma	17
2.4.3 Matriz de Correlação	19
2.4.4 Gráfico de Dispersão (Var.Saída vs Var.Entrada .	22
2.4.5 Variáveis Categóricas	27
2.4.6 Normalização:	29
2.5 Modelação dos dados	30
2.6 Avaliação:	35
3 Resultados e Análise	39

Lista de Figuras

1	Etapas CRISP-DM	3
2	Configurações globais	9
3	Código para a leitura e informações do DataFrame	9
4	Leitura do DataFrame inicialmente	10
5	Estatísticas descritivas	10
6	Leitura do DataFrame depois da limpeza	11
7	Estatísticas descritivas após a limpeza dos dados	12
8	Limpeza dos dados	12
9	Boxplot de Solvent, Order, TPC e ANT	13
10	Boxplot Vm-ratio, Time e AOA	15
11	Histogramas	18
12	Matriz de Correlação	21
13	Gráfico de Dispersão (Var.Saída vs Var.Entrada	23
14	Gráfico de Dispersão (Var.Saída vs Var.Entrada)	28
15	Normalização:	30
16	Código RandomForestRegressor com Normalização	33
17	Valores obtidos com RandomForestRegressor	33
18	Normalização:	36

Resumo

Este trabalho aborda a análise de dados laboratoriais de amostras de beterraba como parte do curso de Análise Inteligente de Dados. Explora-se a importância da análise de dados na extração de compostos bioativos da beterraba, visando não só a otimização dos processos, mas também insights cruciais para as indústrias alimentar e farmacêutica. Utilizaram-se técnicas avançadas de modelagem e estatística para compreender as relações entre variáveis, buscando conclusões acionáveis para avançar na pesquisa nesse campo vital para a nutrição e a saúde.

1 Introdução

No trabalho em questão, está sendo enfrentado o desafio de analisar um conjunto de dados de análises laboratoriais de amostras de beterraba, como parte do curso de Análise Inteligente de Dados. A análise de dados desempenha um papel essencial em diversas áreas do conhecimento, desde a ciência de dados até à estatística aplicada. Seu propósito é extrair insights significativos de conjuntos de dados complexos.

Neste contexto específico, o principal objetivo é extrair informações valiosas dos dados das amostras de beterraba. Pretende-se desenvolver um modelo inteligente capaz de compreender e modelar o processo de análise dessas amostras. Esta modelagem não só melhora a eficiência da extração de compostos bioativos, mas também oferece insights cruciais para a indústria alimentar e farmacêutica.

Ao longo do trabalho, não apenas se explorarão os dados brutos das amostras, mas também serão utilizadas técnicas estatísticas e de modelagem de dados avançadas para compreender melhor as relações entre diferentes variáveis. O objetivo é apresentar conclusões claras e acionáveis que possam informar não apenas a extração de compostos bioativos da beterraba, mas também contribuir para o avanço contínuo da pesquisa neste campo vital para a nutrição e a saúde.

Esta introdução mais detalhada proporciona uma visão mais abrangente do contexto e dos objetivos do estudo, destacando a importância da análise de dados para a compreensão e otimização dos processos relacionados às amostras de beterraba.

2 Metodologia

Na era atual, a capacidade de extrair insights valiosos dos dados é crucial para organizações. O CRISP-DM é uma metodologia usada para esse fim, composta por seis etapas: compreensão do negócio, dos dados, preparação, modelagem, avaliação dos resultados e implementação. Essas etapas fornecem uma estrutura sistemática para análise de dados, visando resultados eficazes.

2.1 Etapas do CRISP-DM

A seguir, serão exploradas as seis etapas do CRISP-DM para uma melhor compreensão de como este modelo foi aplicado no projeto.

1. **Compreensão do Negócio:** Nesta etapa, o objetivo é entender os objetivos do negócio e como a análise de dados pode contribuir para alcançá-los. Isso inclui identificar os problemas de negócio, definir os objetivos do projeto de análise de dados e comunicar-se com os stakeholders para garantir alinhamento entre os objetivos e as necessidades do negócio.

2. **Compreensão dos Dados:** Aqui, busca-se obter uma visão geral dos dados disponíveis e suas características. Isso envolve explorar os dados, identificar padrões e tendências, e validar a qualidade dos dados. A análise exploratória de dados (EDA) é uma técnica comumente utilizada nesta etapa para visualizar e resumir os dados.
3. **Preparação dos Dados:** Esta etapa é crucial, onde os dados brutos são preparados para análise. Isso inclui limpeza de dados para correção de erros, integração de fontes de dados diferentes, seleção de variáveis importantes e transformação para atender aos requisitos do modelo. Garante que os dados estejam prontos para serem utilizados de forma eficaz na análise.
4. **Modelagem dos Dados:**

Nesta etapa, são selecionados e aplicados modelos de mineração de dados aos dados pré-processados, com o objetivo de gerar insights e embasar decisões. Isso pode incluir técnicas de aprendizado supervisionado, não supervisionado ou híbrido, dependendo dos objetivos do projeto.
5. **Avaliação dos Resultados:**

Aqui, avalia-se a precisão, robustez e generalização dos modelos gerados na etapa de Modelagem. Isso envolve a avaliação de métricas de desempenho, como acurácia, precisão, recall e F1-score, bem como a análise de matrizes de confusão e gráficos de ROC.
6. **Implementação:**

Na última etapa, os modelos gerados na etapa de Modelagem são implementados num ambiente de produção, onde podem ser utilizados para gerar insights e embasar decisões. Isso pode incluir a integração com sistemas existentes, a criação de interfaces de utilizador e a implementação de mecanismos de monitorização e avaliação contínua.

Essas etapas fornecem uma estrutura clara para guiar o processo de análise de dados, garantindo que as organizações possam extrair o máximo valor de seus dados de forma eficiente e eficaz.



Figura 1: Etapas CRISP-DM

2.2 Compreensão do problema

A beterraba (*Beta vulgaris*) é conhecida pela sua riqueza em compostos bioativos, como fenólicos, antocianinas e antioxidantes, que apresentam uma série de benefícios para a saúde humana, incluindo propriedades antioxidantes, anti-inflamatórias e cardioprotetoras. Esses compostos têm despertado um interesse crescente devido ao seu potencial terapêutico e nutricional. No entanto, a eficiência da extração desses compostos é crucial para aproveitar plenamente seus benefícios.

2.2.1 Estado da Arte na Análise de Dados de Amostras de Beterraba

Desde os primeiros estudos sobre a extração de compostos bioativos da beterraba, a análise das variáveis experimentais tem sido fundamental para otimizar o processo. A escolha do solvente, da proporção entre o volume e a massa, da ordem de extração e do tempo de extração desempenha um papel crucial na eficácia e na qualidade dos extratos obtidos.

- **Evolução dos Estudos**

Os estudos iniciais exploraram uma variedade de solventes e condições de extração para determinar a sua capacidade de extrair compostos bioativos da beterraba. Com o tempo, esta pesquisa evoluiu para investigar não apenas a eficácia de extração, mas também a segurança e a viabilidade dos solventes em diferentes aplicações, especialmente em alimentos e produtos farmacêuticos.

- **Solventes Utilizados**

Os solventes mais comuns incluem metanol, etanol:água e acetona:água. O metanol, apesar de ser altamente eficaz na extração de compostos bioativos, é limitado em aplicações alimentares devido à sua toxicidade. O etanol:água é amplamente utilizado devido à sua segurança e eficácia, enquanto a acetona:água é eficaz para compostos menos polares.

- **Razão Volume/Massa**

A proporção entre o volume do solvente e a massa da amostra é outro fator crítico. Razões mais baixas, como 5 (25mL/5g), resultam em extratos mais concentrados, enquanto razões mais altas, como 20 (100mL/5g), maximizam o rendimento total, mas podem diluir os compostos extraídos.

- **Ordem de Extração**

A ordem de extração refere-se à sequência em que o solvente é aplicado à amostra. A primeira extração geralmente é mais eficiente em termos de rendimento inicial, enquanto as extrações subsequentes visam recuperar compostos residuais, embora com eficiência decrescente.

- **Tempo de Extração**

O tempo de extração desempenha um papel crucial na eficácia e na qualidade dos extratos. Tempos mais curtos, como 15 minutos, são adequados para compostos mais solúveis e para evitar a degradação, enquanto tempos mais longos, como 60 minutos, permitem uma extração mais completa, mas podem resultar na degradação de compostos sensíveis.

2.3 Compreensão dos dados

Neste capítulo, é importante abordar a descrição dos dados, suas características, variáveis e outros aspectos relevantes. Desta forma, o problema a ser abordado neste estudo é a análise de dados laboratoriais de amostras de beterraba, com o objetivo de extrair insights significativos e desenvolver modelos inteligentes para prever variáveis importantes relacionadas às propriedades das amostras de beterraba. Ao longo do trabalho, seguir-se-ão os passos do CRISP-DM, de forma a obter vários gráficos do estudo em cada etapa. Mais tarde, serão colocados em prática os modelos discutidos no estado da arte para que, depois, seja possível obter os valores do MSE, RMSE, R^2 e MAPE, de forma a entender qual o modelo mais preciso e adequado para ser usado neste estudo da beterraba.

2.3.1 Variáveis de Análise

As análises laboratoriais das amostras de beterraba fornecidas contêm várias variáveis que desempenham papéis fundamentais no processo de análise. Existem dois tipos de variáveis as de saída e de entrada, vamos explorar cada uma delas:

Variáveis de Entrada:

Solvente - O solvente (SOL) é uma substância capaz de dissolver outras substâncias (solutos), formando uma solução. No contexto deste estudo, os solventes são utilizados para extrair componentes específicos das amostras de beterraba. O solvente utilizado nas amostras de beterraba é uma variável categórica com três níveis distintos, onde cada nível representa um tipo diferente de solvente. Os níveis são codificados numericamente como 1, 2 e 3, correspondendo aos seguintes solventes:

- Nível 1: Metanol
- Nível 2: Etanol:Água
- Nível 3: Acetona:Água

Esta codificação permite identificar facilmente qual solvente foi utilizado em cada amostra de beterraba e facilita a análise estatística e o desenvolvimento de modelos preditivos. No contexto do estudo, o tipo de solvente pode influenciar as propriedades extraídas das amostras de beterraba e, consequentemente, pode ser uma variável importante a ser considerada na construção dos modelos preditivos.

Vm-ratio - O (VMR) é razão volume-massa e refere-se à relação entre o volume do solvente utilizado e a massa da amostra de beterraba. Esta variável também é categórica e possui três níveis distintos, codificados numericamente como 5, 10 e 20, correspondendo às seguintes relações:

- Nível 5: 25 mL de solvente por 5 g de amostra
- Nível 10: 50 mL de solvente por 5 g de amostra
- Nível 20: 100 mL de solvente por 5 g de amostra

Essa codificação permite uma diferenciação clara das proporções de volume-massa, facilitando análises precisas. A relação entre volume e massa pode afetar a eficácia da extração dos componentes das amostras, sendo uma variável crucial na construção de modelos preditivos devido às variações na concentração dos componentes extraídos.

Ordem - A variável ordem (ORD) refere-se à sequência das extrações realizadas nas amostras de beterraba. Esta variável é categórica e possui três níveis distintos, codificados numericamente como 1, 2 e 3, representando:

- Nível 1: Primeira extração
- Nível 2: Segunda extração
- Nível 3: Terceira extração

Essa codificação ajuda a identificar claramente a sequência de extrações realizadas, permitindo uma análise detalhada dos dados. A ordem das extrações pode influenciar a quantidade e a qualidade dos componentes extraídos, sendo, portanto, uma variável significativa para a análise estatística e a modelagem preditiva.

Tempo - A variável tempo (TIME) representa a duração do processo de extração das amostras de beterraba, com dois níveis distintos: 15 e 60 minutos. Esta variável é categórica, codificada como 15 e 60, correspondendo aos seguintes tempos de extração:

- Nível 15: 15 minutos de extração
- Nível 60: 60 minutos de extração

O tempo de extração é um fator crucial, pois pode influenciar a eficiência da extração dos componentes das beterrabas. Extrações mais longas podem permitir uma maior dissolução dos componentes, resultando em uma concentração mais elevada dos extratos. Por outro lado, tempos mais curtos podem ser suficientes para certos componentes, economizando tempo e recursos. Portanto, a duração da extração é uma variável importante a ser considerada na análise e modelagem dos dados.

Variáveis de Saída:

TPC - Os Compostos Fenólicos Totais (TPC) referem-se à quantidade global de compostos fenólicos presentes nas amostras de beterraba, reconhecidos pelas suas propriedades antioxidantes. A medição dos TPC é essencial para avaliar a qualidade nutricional dos extratos.

Os níveis de TPC podem variar dependendo do tipo de solvente utilizado, da relação volume-massa (VM-ratio), da sequência das extrações e do tempo de extração. Diferentes solventes e condições de extração, como metanol ou acetona-água, influenciam a quantidade de compostos fenólicos extraídos. Geralmente, uma maior relação volume-massa e tempos de extração mais longos resultam em concentrações mais elevadas de TPC.

A análise dos TPC é crucial para compreender as propriedades antioxidantes e otimizar os processos de extração, visando a obtenção de extratos de alta qualidade.

ANT - As Antocianinas (ANT) são compostos pigmentados naturais encontrados em alimentos como a beterraba, conferindo tonalidades avermelhadas ou arroxeadas. A quantificação das ANT é essencial para avaliar seu conteúdo nas amostras de beterraba, dado o potencial antioxidante desses compostos e seus benefícios para a saúde.

A quantidade de ANT na beterraba pode variar conforme fatores como o cultivo, maturação e processamento do vegetal. Métodos analíticos, como a cromatografia líquida de alta eficiência (HPLC), são utilizados para quantificar com precisão as ANT nesse alimento.

Essa análise é crucial para compreender o potencial antioxidante das ANT presentes na beterraba e seus efeitos na saúde. A presença desses compostos nesse vegetal pode contribuir para a proteção contra danos oxidativos e oferecer benefícios adicionais à saúde, como a redução do risco de doenças cardiovasculares e inflamatórias. Assim, a quantificação precisa das ANT é fundamental para entender e maximizar o valor nutricional da beterraba.

AOA - A Atividade Antioxidante é uma medida da capacidade de uma substância em neutralizar os radicais livres, protegendo as células contra danos oxidativos. A medição da AOA é essencial para avaliar o potencial antioxidante dos extratos de beterraba e sua contribuição para a saúde.

Os níveis de AOA podem variar dependendo de vários fatores, como o tipo de solvente utilizado na extração dos compostos fenólicos, o tempo de extração e a concentração dos antioxidantes presentes. Diferentes métodos de extração,

como a maceração ou a extração por ultrassom, podem influenciar a AOA dos extratos. Geralmente, extratos obtidos com solventes mais eficientes na extração de compostos fenólicos tendem a exibir maior atividade antioxidante.

A análise da AOA é crucial para compreender o potencial antioxidante dos extratos de beterraba e seu impacto na proteção celular contra danos oxidativos. Além disso, a otimização dos processos de extração pode maximizar a AOA dos extratos, aumentando assim seus benefícios para a saúde.

2.3.2 Objetivos do Trabalho:

1. **Extrair o máximo de informação a partir dos dados:** O foco principal é explorar e analisar os dados laboratoriais das amostras de beterraba de maneira abrangente e detalhada. Isso envolve a aplicação de técnicas de análise de dados para identificar padrões, tendências e insights significativos que possam ser úteis na compreensão das características das amostras.
2. **Desenvolver um modelo inteligente capaz de modelar o processo de análise com o melhor desempenho possível:** O objetivo é criar um modelo preditivo robusto e preciso que possa prever variáveis importantes relacionadas às propriedades das amostras de beterraba. Isso envolve a utilização de técnicas avançadas de aprendizado de máquina e análise estatística para desenvolver um modelo que se ajuste bem aos dados e possa fornecer previsões confiáveis.
3. **Apresentar as conclusões com o máximo de clareza possível:** Uma vez concluída a análise dos dados e desenvolvido o modelo preditivo, é essencial apresentar as conclusões de forma clara, concisa e compreensível. Isso inclui a interpretação dos resultados obtidos, a discussão das implicações práticas e teóricas das descobertas e a comunicação eficaz das conclusões para os interessados no estudo.

2.3.3 Código usado no Python nesta etapa:

É relevante destacar que, anteriormente, foram abordadas duas etapas do CRISP-DM de forma teórica. No entanto, em termos práticos, consideramos que apenas a partir da fase de Entendimento dos Dados é apropriado começar a aplicar código no Spyder. Isso ocorre porque a compreensão do problema é fundamental antes de começarmos a trabalhar com o código. É necessário entender completamente o problema e os dados disponíveis antes de começar a manipulá-los e analisá-los no ambiente do Spyder.

- Excell to Python

Antes de iniciar formalmente as etapas do CRISP-DM, é essencial converter os dados fornecidos de um formato Excel para um DataFrame no Spyder. Isso permite o acesso aos dados e é uma etapa crucial para preparar a análise subsequente.

```
# In[1] Configurações globais  
  
# Colocar o ficheiro em Excell  
sns.set()  
excel_path = r"C:\Users\Andreia\Desktop\beetroot\Data\Beetroot.xlsx"
```

Figura 2: Configurações globais

- Leitura e informações sobre o DataFrame

Após o DataFrame ter sido importado com sucesso do Excel para o Python, torna-se possível escrever um código que efetue a leitura do DataFrame e forneça informações sobre o mesmo antes de serem realizadas quaisquer alterações nos dados. Isso ajuda a compreender a estrutura e o conteúdo do DataFrame inicial, facilitando a identificação de padrões e características importantes para análises posteriores.

```
# In[2] Data Understanding:  
  
# Ler o DataFrame  
df = pd.read_excel(excel_path);  
  
# Obter informações sobre o DataFrame  
print(df.info());  
  
# Obter Dados do DataFrame  
print("\n\nColumns: ");  
for column in df.columns:  
    print(column);  
  
print(df.head());  
  
print("\n Estatísticas Descritivas do DataFrame:")  
print("\n", df.describe());
```

Figura 3: Código para a leitura e informações do DataFrame

Após a execução do código, é possível visualizar no Python o DataFrame, juntamente com suas informações e dados associados. Isso proporciona uma compreensão mais clara da estrutura e do conteúdo do DataFrame, o que facilita a análise e a identificação de padrões ou características importantes para o processo de mineração de dados. Abaixo, nas figuras, podemos ver duas partes do DataFrame que serão usadas como exemplos nos próximos processos.

Figura 4: Leitura do DataFrame inicialmente

Por fim, ainda foi possível obtermos diferentes valores das várias variáveis, tanto de entrada como saída, estes valores são estatísticas descritivas das colunas do DataFrame, que fornecem informações como contagem, média, desvio padrão, mínimo, máximo e quartis, como podemos verificar na figura abaixo representada .

Descrição Estatística:							
	Solvent	Vm-ratio	Order	...	TPC	AOA	ANT
count	189.000000	189.000000	189.000000	...	189.000000	182.000000	189.000000
mean	1.857143	11.656667	2.000000	...	0.878190	6.432418	0.198847
std	0.835206	6.252659	0.818665	...	0.494416	6.338255	0.226180
min	1.000000	5.000000	1.000000	...	0.127000	0.000000	0.000000
25%	1.000000	5.000000	1.000000	...	0.500000	1.400000	0.020000
50%	2.000000	10.000000	2.000000	...	0.730000	3.400000	0.080000
75%	3.000000	20.000000	3.000000	...	1.247000	13.075000	0.418000
max	3.000000	20.000000	3.000000	...	2.090000	22.400000	0.770000

Figura 5: Estatísticas descritivas

As estatísticas revelam insights sobre diferentes variáveis. O solvente mostra uma distribuição equilibrada, enquanto Vm-ratio tem uma variação considerável. As ordens são uniformemente distribuídas. TPC tem uma média de 0.878, com a maioria dos valores perto da mediana de 0.730. AOA tem alta variabilidade e presença de outliers, enquanto ANT mostra menor variabilidade. Valores ausentes em AOA indicam a necessidade de tratamento. Essas observações são cruciais para entender os dados e preparar análises futuras.

2.4 Preparação dos dados

Nesta fase, de preparação dos dados (Data Understanding), continuamos a análise, visando descrever detalhadamente os dados e prepará-los para a modelagem. O objetivo é assegurar que os dados estejam prontos e adequados para análises estatísticas e modelagem, permitindo-nos extrair insights significativos e desenvolver modelos inteligentes para prever variáveis importantes relacionadas às propriedades das amostras de beterraba.

Durante esta fase, é essencial limpar o DataFrame para garantir a sua qualidade. Conforme observado anteriormente na Figura 4, existem vários valores NaN, duplicados, zeros e outliers. Portanto, o objetivo é remover os valores NaN e duplicados, além de substituir os valores zero pela média por coluna e, por fim, remover os valores discrepantes (como se pode observar na figura abaixo). Estas ações são fundamentais para garantir que os dados estejam prontos para a modelagem e análise subsequente.

index	latitude	longitude	depth	time	temp	pH	sal	dist
1	5	5	1	15	0.89	11.2	0.43	
2	5	5	1	15	0.91	10.5	0.42	
3	5	5	1	15	0.91	11.1	0.43	
4	5	5	1	15	0.9	10.3	0.39	
5	5	5	1	15	0.9	11.1	0.39	
6	5	5	1	15	0.91	11.1	0.39	
7	5	5	2	15	0.46	4.4	0.01	
8	5	5	2	15	0.46	5.3	0.1	
9	5	5	2	15	0.46	4.4	0.15	
10	5	5	2	15	0.46	5.3	0.14	
11	5	5	2	15	0.47	4.6	0.15	
12	5	5	2	15	0.47	4.6	0.15	
13	5	5	3	15	0.43	0.6	0.230019	
14	5	5	3	15	0.5	0.8	0.230019	
15	5	5	3	15	0.36	0.8	0.05	
16	5	5	3	15	0.3	0.9	0.04	
17	5	5	3	15	0.41	1.3	0.03	
18	5	10	1	15	1.1	16.7	0.47	
19	5	10	1	15	1.89	16.7	0.58	
20	5	10	1	15	1.18	15.3	0.48	
21	5	10	1	15	1.87	12.8	0.54	
22	5	10	1	15	1.89	15.2	0.43	
23	5	10	1	15	1.17	15.2	0.53	
24	5	10	2	15	0.96	4.1	0.15	
25	5	10	2	15	0.98	5.4	0.17	
26	5	10	2	15	0.63	4.1	0.15	
27	5	10	2	15	0.61	4.1	0.16	
28	5	10	2	15	0.61	4.5	0.15	
29	5	10	3	15	0.45	0.8	0.07	
30	5	10	3	15	0.62	0.8	0.07	
31	5	10	3	15	0.59	0.8	0.230019	

index	latitude	longitude	depth	time	temp	pH	sal	dist
90	5	5	3	15	0.33	0.3	0.08	
91	5	10	1	15	1.43	16.3	0.23	
92	5	10	1	15	1.37	15.5	0.23	
93	5	10	1	15	1.43	16	0.22	
94	5	10	2	15	0.73	2.4	0.230019	
95	5	10	2	15	0.73	2.7	0.230019	
96	5	10	2	15	0.77	2.4	0.230019	
97	5	10	3	15	0.54	0.2	0.230019	
98	5	10	3	15	0.54	7.01018	0.230019	
99	5	10	3	15	0.58	7.01018	0.230019	
100	5	20	1	15	2.46	18.5	0.57	
101	5	20	1	15	2.41	18.8	0.62	
102	5	20	2	15	1.32	3.3	0.05	
103	5	20	2	15	1.43	3.7	0.05	
104	5	20	2	15	1.28	3.8	0.06	
105	5	20	3	15	0.79	7.01018	0.230019	
106	5	20	3	15	0.69	7.01018	0.230019	
107	5	20	3	15	0.71	7.01018	0.230019	
108	5	5	2	00	0.7	13.6	0.11	
109	5	5	2	00	0.73	12.7	0.11	
110	5	5	2	00	0.72	13.6	0.18	
111	5	5	2	00	0.34	3.2	0.05	
112	5	5	2	00	0.35	3.2	0.05	
113	5	5	2	00	0.35	3.4	0.05	
114	5	5	3	00	0.22	2.1	0.03	
115	5	5	3	00	0.21	1.8	0.02	
116	5	5	3	00	0.24	1.9	0.03	
117	5	10	1	00	1.81	14.2	0.44	
118	5	10	1	00	1.89	15.5	0.42	
119	5	10	1	00	1.12	15	0.42	

Figura 6: Leitura do DataFrame depois da limpeza

Após a limpeza dos dados, observamos uma diminuição nas estatísticas descritivas, como média, desvio padrão, mínimo, máximo e quartis. Isso ocorre devido à remoção de outliers, que podem distorcer as medidas de tendência central e dispersão dos dados. Ao remover os valores extremos, as estatísticas tornam-se mais representativas da distribuição subjacente, resultando em uma média mais precisa, menor dispersão e uma distribuição mais homogênea entre os quartis.

	Solvent	Vm-ratio	Order	...	TPC	AOA	ANT
count	178.000000	178.000000	178.000000	...	178.000000	178.000000	178.000000
mean	1.870787	11.432584	2.011236	...	0.852764	6.808723	0.231265
std	0.830316	6.179834	0.823310	...	0.468893	5.943562	0.207134
min	1.000000	5.000000	1.000000	...	0.127000	0.200000	0.010000
25%	1.000000	5.000000	1.000000	...	0.500000	2.000000	0.060000
50%	2.000000	10.000000	2.000000	...	0.728500	4.000000	0.155000
75%	3.000000	20.000000	3.000000	...	1.185000	12.950000	0.407500
max	3.000000	20.000000	3.000000	...	2.050000	22.400000	0.770000

Figura 7: Estatísticas descritivas após a limpeza dos dados

Após a remoção de outliers, as estatísticas revelam ajustes nas médias e nos desvios padrão das variáveis. Em geral, as distribuições permanecem similares às originais, porém mais precisas. Notavelmente, houve redução na variação extrema, resultando em distribuições mais uniformes e médias mais confiáveis. Este refinamento dos dados promove uma análise mais robusta e confiável para futuras investigações.

O código utilizado para realizar todas essas alterações e calcular os valores médios antes da limpeza do DataFrame foi o seguinte:

```
# In[3] Data Preparation:

# Remover linhas com valores NaN
df = df.dropna()

# Valores duplicados
df = df.drop_duplicates()

# Substituir os zeros pela média das colunas
for coluna in df.columns:
    media_coluna = df[coluna][df[coluna] != 0].mean()
    df[coluna] = df[coluna].replace(0, media_coluna)

# Identificar e remover outliers
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
outliers = ((df < (Q1 - 1.1 * IQR)) | (df > (Q3 + 1.1 * IQR))).any(axis=1)
df_sem_outliers = df[~outliers]

# Exibir estatísticas dos dados sem outliers
print("\nEstatísticas sem outliers:")
print(df_sem_outliers.describe())
```

Figura 8: Limpeza dos dados

Após limpar os dados, empregamos diversas técnicas de visualização, como boxplots e histogramas, para compreender a distribuição e relação entre as variáveis. Observamos a dispersão das variáveis em relação às saídas TPC, AOA e ANT e identificamos padrões e correlações por meio de gráficos de dispersão. Essas análises visuais forneceram insights cruciais para orientar a modelagem e análise posteriores.

2.4.1 Boxplot

O boxplot é uma representação visual compacta e informativa da distribuição de um conjunto de dados numéricos. Ele apresenta cinco estatísticas resumidas:

mínimo, primeiro quartil (Q1), mediana, terceiro quartil (Q3) e máximo. Além disso, identifica outliers, que são valores fora do intervalo interquartil. Essa ferramenta é útil para identificar a dispersão dos dados, detectar outliers e comparar distribuições entre diferentes grupos de dados.

É importante fornecer os valores destas variáveis mencionadas acima, desta forma, também podemos perceber que os valores dados inicialmente da Data-frame antes de qualquer alteração irá ser diferente dos valores fornecidos abaixo.

O boxplot 1 mostra a distribuição da variável Solvent, Order, TPC e ANT.

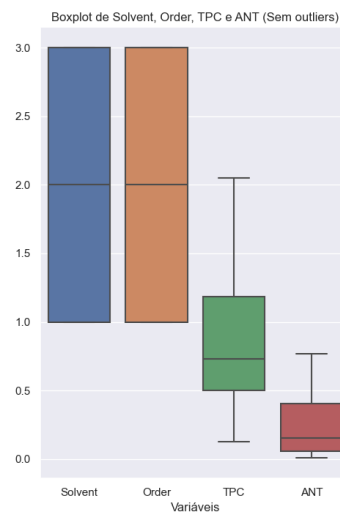


Figura 9: Boxplot de Solvent, Order, TPC e ANT

Ao analisar o boxplot obtido para as variáveis Solvent, Order, TPC e ANT chegamos às seguintes conclusões:

Solvent e Order

Boxplot: Ambos têm a mesma distribuição e valores.

Valores:

- Mínimo: 1
- Primeiro quartil (Q1): 1
- Mediana: 2

- Terceiro quartil (Q3): 3
- Máximo: 3

As variáveis solvent e order são categóricas e compreendem valores inteiros de 1 a 3. Após a análise é possível perceber que a maioria dos dados se encontra nos extremos (1 e 3), com uma mediana de 2.

TPC

Boxplot: A caixa estende-se do primeiro quartil ($Q1 = 0.5$) ao terceiro quartil ($Q3 = 1.185$). A linha no meio da caixa representa a mediana (0.7285). Os bigodes da caixa alogam-se até ao mínimo (0.127) e até ao máximo (2.05), e representam valores que não ocorrem com abundância, porque não está inserido na caixa. O gráfico é representado sem outliers visto que o código de criação do gráfico foi corrido sem os mesmos.

Valores:

- Mínimo: 0.127
- Primeiro quartil (Q1): 0.5
- Mediana: 0.7285
- Terceiro quartil (Q3): 1.185
- Máximo: 2.05

A maioria dos valores de TPC está concentrada entre 0.5 e 1.185, representados na caixa, com a mediana ligeiramente afastada do centro, encontra-se mais para baixo, o que demonstra uma leve assimetria. Os seus bigodes são relativamente curtos o que indica que não há valores extremos muito altos ou muito baixos.

ANT

Boxplot: A caixa desta variável vai desde o primeiro quartil, de valor igual a 0.06, até ao terceiro quartil, de valor igual a 0.4075. A mediana tem valor igual a 0.155 e os bigodes têm um máximo de 0.77 e um mínimo de 0.01, sem outliers visíveis.

Valores:

- Mínimo: 0.01
- Primeiro quartil (Q1): 0.06
- Mediana: 0.155
- Terceiro quartil (Q3): 0.4075
- Máximo: 0.77

A maioria dos valores de ANT está concentrada entre 0.06 e 0.4075, com a mediana mais próxima ao valor mais baixo, indicando uma distribuição assimétrica. Os bigodes são relativamente curtos, sugerindo que a distribuição dos dados é compacta com poucos valores extremos.

As variáveis Solvent e Order apresentam uma distribuição categórica com valores de 1 a 3, com uma distribuição uniforme entre as categorias. As variáveis TPC e ANT têm distribuições contínuas com diferentes graus de assimetria e dispersão. A ausência dos outliers demonstra que os dados foram limpos e filtrados, como esperado.

Estes boxplots são úteis para comparar as distribuições das variáveis e identificar rapidamente a mediana, a dispersão e a presença de assimetrias nos dados.

O boxplot 2 mostra a distribuição da variável AOA em relação às variáveis vm-ratio e time.

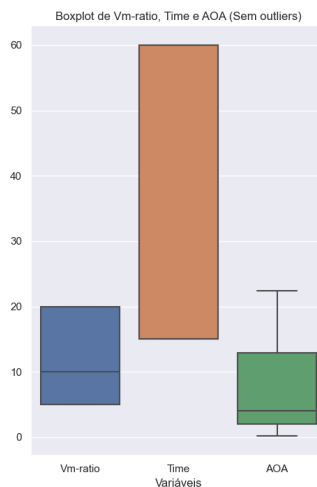


Figura 10: Boxplot Vm-ratio, Time e AOA

Ao analisar o boxplot com as variáveis Vm-ratio, Time e AOA podemos tirar algumas conclusões como:

Vm-ratio:

Boxplot: A caixa estende-se do primeiro quartil (5.0) até ao terceiro quartil (20.0). Tem uma mediana de valor 10.0 e não contém bigodes visto que os extremos compreendem os dados com maior abundância, coincidindo com os quartis.

Valores:

- Mínimo: 5.0
- Primeiro Quartil (Q1): 5.0
- Mediana (Q2): 10.0
- Terceiro Quartil (Q3): 20.0
- Máximo: 20.0

A distribuição dos valores de Vm-ratio é bastante concentrada em torno dos valores mínimos e máximos permitidos. A mediana está bem no centro da distribuição, indicando que metade dos dados estão abaixo de 10.0 e a outra metade acima.

Time: Vm-ratio:

Boxplot: A caixa estende-se do primeiro quartil (15.0) até ao terceiro quartil (60.0). A mediana está em 15.0, coincidente com o primeiro quartil. Não contém bigodes visto que os extremos compreendem os dados com maior abundância, coincidindo com os quartis.

Valores:

- Mínimo: 15.0
- Primeiro Quartil (Q1): 15.0
- Mediana (Q2): 15.0
- Terceiro Quartil (Q3): 60.0
- Máximo: 60.0

A distribuição dos valores de Time mostra que metade dos dados estão no valor mínimo de 15.0, indicando uma forte concentração inicial. O restante dos dados se distribui até 60.0, mostrando uma alta variabilidade com valores dispersos até o máximo.

AOA:

Boxplot: A caixa estende-se do primeiro quartil (2.0) até o terceiro quartil (12.95) e a mediana tem um valor de 4.0. O bigode inferior vai desde o valor mínimo 0.2 até ao valor máximo de 22.4.

Valores:

- Mínimo: 0.2
- Primeiro Quartil (Q1): 2.0
- Mediana (Q2): 4.0
- Terceiro Quartil (Q3): 12.95
- Máximo: 22.4

A distribuição dos valores de AOA mostra uma maior concentração de valores na parte inferior ao interquartil, com a mediana mais próxima do primeiro quartil. A ampla extensão dos bigodes indica a presença de uma alta variabilidade nos dados, com valores que se dispersam até o máximo.

A interpretação deste boxplot ajuda a entender a dispersão, a centralização e a variabilidade dos dados, facilitando a análise de tendências e possíveis anomalias nas distribuições. Assim, após a observação do boxplot, percebemos que a variável Vm-ratio demonstra uma distribuição concentrada entre os valores 5.0 e 20.0, com a mediana centralizada, a variável Time possui uma alta concentração no valor mínimo (15.0) com dispersão até o máximo (60.0) e a variável AOA apresenta uma variabilidade significativa com valores concentrados nos primeiros quartis e dispersão ampla até o valor máximo.

2.4.2 Histograma

Um histograma é uma representação gráfica que mostra a distribuição de frequências de um conjunto de dados numéricos. É útil para mostrar a distribuição das amostras utilizando o parâmetro bins, que indica o número de classes. Cria uma visualização que permite entender a forma e a densidade da distribuição dos dados.

Os histogramas são ferramentas fundamentais para diversas análises, como a visualização da distribuição de dados, identificação de tendências e padrões, comparação de distribuições de diferentes amostras e detecção de outliers. A facilidade de interpretação é uma das suas maiores vantagens, pois oferece uma representação visual simples e intuitiva da distribuição dos dados, facilitando a comunicação de informações complexas de maneira clara e concisa.

A flexibilidade dos histogramas, que permite ajustar o número de intervalos para uma análise mais detalhada, torna-os uma escolha popular em várias áreas de estudo e aplicação. Em resumo, histogramas são essenciais na análise de dados, proporcionando insights valiosos sobre a distribuição e o comportamento dos dados de uma forma visualmente acessível e compreensível.

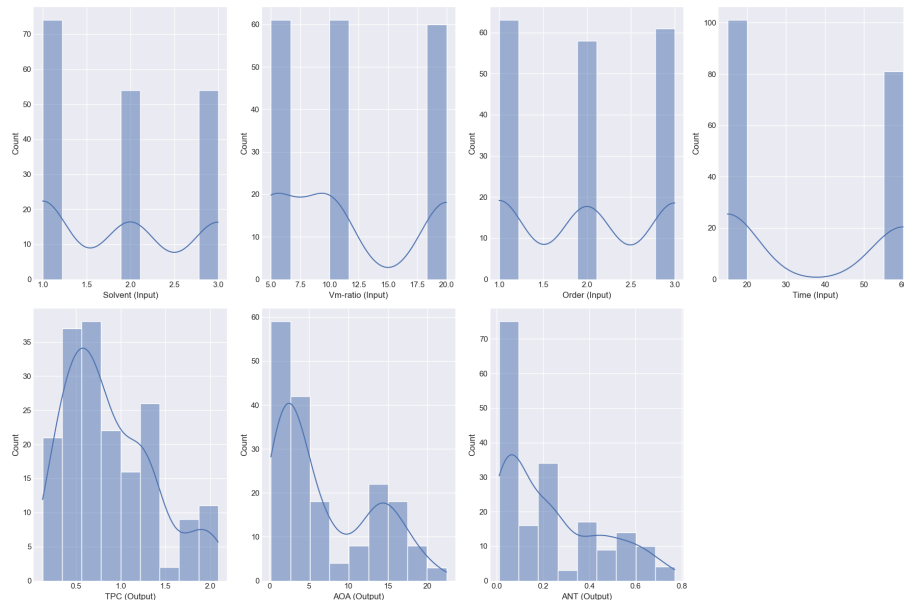


Figura 11: Histogramas

Na imagem acima apresentada observamos o histograma obtido através do código para a análise. A figura mostra seis histogramas que representam a distribuição de diferentes variáveis (tanto de entrada como de saída) da amostra de beterraba. Cada histograma é acompanhado por uma linha de densidade que ajuda a visualizar a forma da distribuição dos dados.

Vamos proceder primeiramente a uma análise das variáveis de entrada (Solvent, Vm-ratio e Ordem): na análise da contagem da experiência do solvent é possível concluir que o valor do solvent 1 é diferente do solvent 2 e do solvent

3, apresenta 3 picos diferentes, indicando uma distribuição multimodal, assim, também conseguimos observar diferentes frequências, sendo a maior igual a 70, que representa o solvent 1; a variável "Vm-ratio" também mostra três picos distintos em torno dos valores 5.0, 10.0 e 20.0, indicando também uma distribuição multimodal, os valores 5.0 e 10.0 têm frequências praticamente iguais, apenas o valor 20.0 difere um pouco mais; os picos da variável "order" são três, nos valores 1.0, 2.0 e 3.0, relativamente às frequências é de notar que o valor em 2.0 é menor que o valor em 3.0, isto pode indicar uma perda de dados, visto que as amostras seguintes precisam sempre de usufruir das amostras anteriores.

Na análise dos gráficos das variáveis de saída podemos retirar também várias informações. Relativamente à variável "TPC" é notória uma distribuição assimétrica mais à direita e, em relação à frequência, a maior frequência é observada em torno de 0.5 a 1.0, com mais de 35 ocorrências. A variável "AOA" apresenta uma distribuição bimodal, mostrando dois picos principais, um entre 0 e 5 e outro entre 15 e 20, a maior frequência está próxima de 0, com cerca de 60 ocorrências. Relativamente à variável "ANT", esta apresenta uma distribuição assimétrica como é possível observar na figura, a maior frequência é observada em valores próximos de 0.0 a 0.1, com mais de 70 ocorrências, enquanto que as restantes amostras têm frequências significativamente mais baixas.

Numa análise geral podemos então concluir que há maioritariamente distribuições multimodais e assimétricas e que as frequências mais altas nas variáveis de entrada indicam que certos valores são mais comuns na amostra enquanto que nas variáveis de saída, a maior concentração em valores mais baixos sugere que a maioria das amostras tem conteúdos relativamente baixos de compostos específicos.

A distribuição das variáveis de entrada pode sugerir diferentes tratamentos ou condições aplicadas às amostras de beterraba, enquanto que as distribuições das variáveis de saída indicam a resposta das amostras a essas condições. Os picos nas distribuições podem ser úteis para identificar grupos ou categorias distintas dentro dos dados, como diferentes métodos de processamento ou categorias de amostras. Compreender estas distribuições pode ajudar a ajustar os processos de produção para otimizar as propriedades desejadas, como a atividade antioxidante.

2.4.3 Matriz de Correlação

A matriz de correlação é uma ferramenta estatística que apresenta os coeficientes de correlação entre pares de variáveis. Estes coeficientes variam de -1 a 1 e indicam a força (com o auxílio do heatmap) e a direção da relação linear entre duas variáveis. Aqui estão os principais valores de referência:

- **1**: Correlação positiva perfeita (quando uma variável aumenta, a outra também aumenta de forma proporcional).
- **-1**: Correlação negativa perfeita (quando uma variável aumenta, a outra diminui de forma proporcional).
- **0**: Nenhuma correlação linear (não há relação linear entre as variáveis).

Heatmap

Um heatmap é um tipo de gráfico que utiliza uma paleta de cores para representar a magnitude dos valores em uma matriz bidimensional. No contexto de uma matriz de correlação, cada célula do heatmap mostra o coeficiente de correlação entre duas variáveis, onde:

- Cores mais próximas do vermelho indicam correlações positivas fortes.
- Cores mais próximas do azul indicam correlações negativas fortes.
- Cores próximas do branco ou neutras indicam correlações fracas ou inexistentes.

Explicação técnica do gráfico de Correlação

A matriz de correlação é uma tabela que exibe os coeficientes de correlação entre pares de variáveis. O coeficiente de correlação varia de -1 a 1, onde:

- **1** indica uma correlação positiva perfeita (à medida que uma variável aumenta, a outra também aumenta).
- **-1** indica uma correlação negativa perfeita (à medida que uma variável aumenta, a outra diminui).
- **0** indica nenhuma correlação linear (as variáveis não têm relação linear entre si).

Conclusões da Matriz de Correlação

Com base na imagem da matriz de correlação apresentada em baixo e os respectivos valores da matriz, podemos tirar várias conclusões:

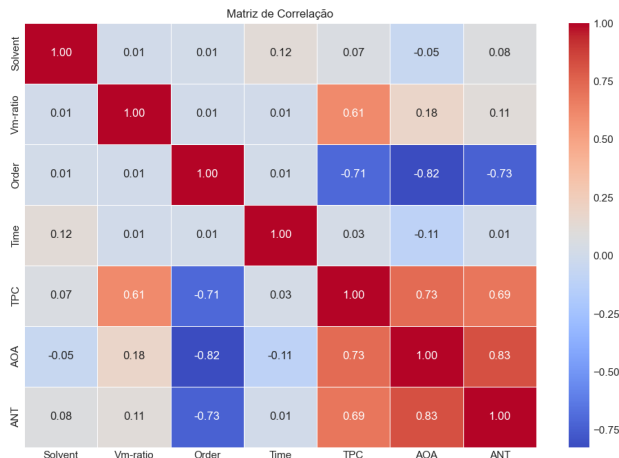


Figura 12: Matriz de Correlação

Correlação Positiva Forte

- **TPC e AOA (0.73):** Existe uma correlação positiva forte entre TPC (Total Phenolic Content) e AOA (Antioxidant Activity), sugerindo que um aumento no conteúdo fenólico total está associado a um aumento na atividade antioxidante.
- **TPC e ANT (0.69):** Também há uma correlação positiva considerável entre TPC e ANT (Anthocyanins).
- **AOA e ANT (0.83):** A atividade antioxidante e as antocianinas têm uma correlação positiva muito forte, indicando que altos níveis de antocianinas estão fortemente associados a alta atividade antioxidante.

Correlação Negativa Forte

- **Order e TPC (-0.71):** Existe uma correlação negativa forte entre Order e TPC, sugerindo que um aumento na ordem está associado a uma diminuição no conteúdo fenólico total.
- **Order e AOA (-0.82):** A ordem também tem uma correlação negativa forte com a atividade antioxidante, similar ao conteúdo fenólico total.

- **Order e ANT (-0.73):** De maneira semelhante, há uma correlação negativa forte entre Order e ANT, indicando que uma ordem maior está associada a níveis menores de antocianinas.

Correlação Positiva Moderada

- **Vm-ratio e TPC (0.61):** A razão Vm e TPC têm uma correlação positiva moderada.

Interpretação

A análise das relações entre TPC, AOA e ANT pode revelar insights importantes sobre as propriedades químicas e biológicas dos compostos estudados. Uma forte correlação positiva sugere que aumentar o conteúdo fenólico pode aumentar a atividade antioxidante e os níveis de antocianinas, úteis em aplicações alimentares e farmacêuticas.

Por outro lado, uma correlação negativa com a variável Order indica que condições associadas ao aumento da ordem podem reduzir esses componentes benéficos. A análise da matriz de correlação por meio de um heatmap oferece uma visão clara das inter-relações entre as variáveis, auxiliando na tomada de decisões informadas em pesquisas e aplicações práticas.

2.4.4 Gráfico de Dispersão (Var.Saída vs Var.Entrada)

A análise dos gráficos de dispersão fornece uma visão detalhada das relações entre as variáveis de entrada e saída no processo de extração de compostos bioativos. É importante tirar conclusões individuais para cada comparação de entrada com saída e identificar o nível mais eficiente dentro de cada variável de entrada. Essa abordagem permite determinar as melhores condições de extração com base nas variáveis fornecidas.

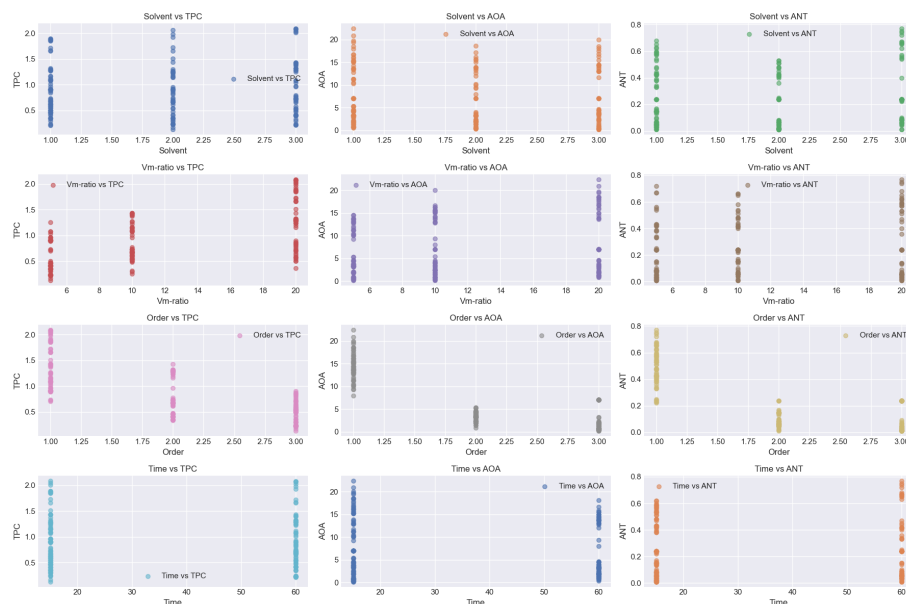


Figura 13: Gráfico de Dispersão (Var.Saída vs Var.Entrada)

- **Solvent vs TPC**

Melhor Solvente: Acetone:Water (Solvente 3).

A variação do solvente impacta os compostos fenólicos totais (TPC), com a Acetone:Water (Solvente 3) gerando as maiores concentrações, apesar de alguns outliers. O Methanol (Solvente 1) produz valores menores. Entre eles, a Acetone:Water é preferível pela sua maior eficácia, apesar dos outliers. O Ethanol é descartado devido às concentrações iniciais baixas.

- **Solvent vs AOA:**

Melhor Solvente: Methanol (Solvente 1).

A atividade antioxidante é influenciada pelos solventes, com Methanol (1) e Acetone:Water (3) apresentando valores superiores em comparação com Ethanol:Water (2). O Methanol parece ser mais eficaz, iniciando sua atividade após o zero, indicando uma possível persistência antioxidante.

- **Solvent vs ANT:**

Melhor Solvente: Acetone:Water (Solvente 3).

O solvente 3 (Acetone:Water) produz a maior quantidade de antocianinas (ANT), embora apresente alguns outliers. O solvente 1 (Methanol)

também atinge valores altos, mas menores que o Acetone:Water e com menos outliers. O solvente 2 (Ethanol:Water) produz valores baixos, sendo irrelevante. Assim, o Acetone:Water é a melhor opção para extração de antocianinas, devido à sua capacidade de alcançar maiores concentrações, mesmo com outliers.

Conclusões Globais Solvent Vs TPC, AOA, ANT:

Em resumo, observamos que a Acetone:Water (Solvente 3) se destaca como a melhor opção em relação ao TPC e ANT, enquanto o Methanol (Solvente 1) é preferível para a AOA. A Acetone:Water (Solvente 3) é possivelmente a melhor opção devido à sua capacidade de produzir as maiores concentrações dos compostos de interesse, apesar de alguns outliers. A Acetona é conhecida por ser um solvente eficaz na extração de compostos fenólicos, incluindo antocianinas, devido à sua polaridade intermediária e capacidade de solvatação adequada. Essa combinação de eficiência na extração e capacidade de solvatação faz da Acetone:Water uma escolha vantajosa para maximizar a quantidade de compostos fenólicos totais e antocianinas obtidos.

- **Vm-ratio vs TPC:**

A razão volume-massa (Vm-ratio) influencia os compostos fenólicos totais (TPC), onde uma maior Vm-ratio (20) está associada a maiores concentrações de TPC. Portanto, podemos concluir que o melhor volume-massa a ser usado é o do nível 20, que corresponde a 100mL/5g.

- **Vm-ratio vs AOA:**

A razão volume-massa (Vm-ratio) influencia a atividade antioxidante (AOA), onde uma maior Vm-ratio (20) está associada a maiores concentrações de AOA. Portanto, podemos concluir que o melhor volume-massa a ser usado é o do nível 20, que corresponde a 100mL/5g.

- **Vm-ratio vs ANT:** Na análise da razão volume-massa (Vm-ratio) em relação às antocianinas (ANT), observamos que uma maior Vm-ratio (20) está correlacionada com maiores concentrações de ANT. Portanto, concluímos que o melhor volume-massa a ser utilizado é o do NÍVEL 20, que corresponde a 100mL/5g. É importante destacar que nesta comparação, os valores dos três níveis são mais uniformes, não apresentando grandes discrepâncias entre si.

Conclusões Globais Vm-Ratio Vs TPC, AOA, ANT:

O valor do Vm-ratio ser o melhor em todas as análises sugere que a proporção de 100mL de solvente para 5g de amostra é ideal para extrair eficazmente compostos fenólicos totais, atividade antioxidante e antocianinas. Isso pode dever-se ao facto de este Vm-ratio proporcionar uma relação otimizada entre a quantidade de solvente disponível para extrair os compostos da amostra e a quantidade de amostra disponível para ser extraída. Esta proporção equilibrada pode garantir uma interação eficaz entre o solvente e os compostos bioativos presentes na amostra, resultando em maiores rendimentos de extração..

- **Order vs TPC:**

É comum observar uma diminuição nos compostos fenólicos totais (TPC) à medida que a ordem de extração aumenta. Isso ocorre porque, durante o processo de extração, os compostos fenólicos presentes na amostra são gradualmente transferidos para o solvente. Na primeira extração (Order 1), a amostra ainda contém uma quantidade significativa de compostos fenólicos, resultando em valores mais altos de TPC. À medida que as extrações subsequentes são realizadas (Order 2 e Order 3), a quantidade de compostos fenólicos remanescentes na amostra diminui, levando a valores menores de TPC. Esse fenómeno é esperado e ocorre devido à exaustão dos compostos fenólicos na matriz da amostra durante as extrações sucessivas.

- **Order vs AOA:**

A diminuição da atividade antioxidante (AOA) com a ordem de extração é um fenómeno observado comumente. Isso ocorre devido à natureza dos compostos antioxidantes presentes na amostra, que são gradualmente transferidos para o solvente durante o processo de extração. Na primeira extração (Order 1), a amostra tende a conter uma concentração mais elevada desses compostos antioxidantes, resultando em valores mais altos de AOA. À medida que as extrações subsequentes são realizadas (Order 2 e Order 3), a quantidade de compostos antioxidantes remanescentes na amostra diminui, levando a uma redução na atividade antioxidante medida. Esse declínio na AOA ao longo das extrações pode ser explicado pela exaustão dos compostos antioxidantes disponíveis na amostra, à medida que são transferidos para o solvente durante o processo de extração.

- **Order vs ANT:**

A quantidade de antocianinas (ANT) é mais alta na primeira extração (Order 1) e diminui nas subsequentes (Order 2 e Order 3). Esse padrão é observado devido à natureza dos compostos de antocianinas na amostra e ao processo de extração. Na primeira extração, a amostra contém uma concentração mais elevada de antocianinas, resultando em valores mais altos de ANT. À medida que as extrações subsequentes são realizadas (Order 2 e Order 3), a quantidade de antocianinas remanescentes na amostra diminui, levando a valores menores de ANT. Esse declínio

na quantidade de antocianinas ao longo das extrações é devido à transferência desses compostos para o solvente durante o processo de extração, esgotando gradualmente a quantidade disponível na matriz da amostra.

Conclusões Globais Order Vs TPC, AOA, ANT:

A ordem de extração desempenha um papel crucial na eficácia dos processos de extração de compostos bioativos. Na primeira extração, há uma eficiência maior na obtenção de compostos fenólicos, antioxidantes e antocianinas, devido à alta concentração inicial desses compostos na amostra. À medida que as extrações subsequentes são realizadas, a quantidade de compostos bioativos na amostra diminui gradualmente. Essas conclusões destacam a importância de um planejamento cuidadoso das etapas de extração e da interpretação adequada dos resultados para garantir a eficiência e precisão na obtenção de compostos bioativos.

- **Time vs TPC:** Os gráficos de dispersão que relacionam o tempo de extração com os compostos fenólicos totais (TPC) mostram uma variação semelhante nos valores de TPC para diferentes tempos de extração, como observado nos tempos de 15 minutos e 60 minutos. Isso pode ser devido a uma extração incompleta devido a um tempo insuficiente, ao equilíbrio dinâmico do processo de extração ou à manutenção de outros fatores externos constantes durante os diferentes tempos de extração. Essa uniformidade destaca a necessidade de considerar outros parâmetros de extração além do tempo para avaliar adequadamente a eficiência da extração de compostos fenólicos.

- **Time vs AOA:**

Embora ambos os tempos de extração mostrem uma ampla distribuição de valores de atividade antioxidante (AOA), uma análise mais detalhada revela uma tendência interessante: o tempo de 15 minutos tende a alcançar valores mais altos de AOA em comparação com o tempo de 60 minutos, onde a AOA diminui ligeiramente. Isso sugere que um tempo de extração mais curto pode resultar em uma extração mais eficiente de compostos com atividade antioxidante. No entanto, a distribuição variada de valores de AOA para ambos os tempos indica que outros fatores além do tempo de extração também podem influenciar a atividade antioxidante dos extratos.

- **Time vs ANT**

Apesar de uma ampla variação nos valores de antocianinas (ANT) para ambos os tempos de extração (15 minutos e 60 minutos), uma observação notável é que os valores de ANT para o tempo de 60 minutos tendem a ser maiores do que para o tempo de 15 minutos. Isso sugere que um tempo de extração mais longo pode resultar em uma extração mais eficaz de antocianinas. No entanto, é importante notar que ainda há uma considerável variação nos valores de ANT para ambos os tempos de extração, indicando

que outros fatores podem estar influenciando a extração de antocianinas além do tempo.

Conclusões Globais Time Vs TPC, AOA, ANT:

Com base nas análises dos gráficos de dispersão, podemos concluir que a escolha entre os tempos de 15 e 60 minutos depende do composto de interesse e dos objetivos da extração. Se o foco principal for a atividade antioxidante, o tempo de 15 minutos parece resultar em valores mais altos de AOA. No entanto, se a prioridade for a extração de antocianinas, o tempo de 60 minutos tende a ser mais eficaz. Portanto, a decisão entre os dois tempos deve levar em consideração o composto alvo e os resultados desejados da extração.

2.4.5 Variáveis Categóricas

As variáveis categóricas são fundamentais na análise estatística e modelagem de dados, representando características organizadas em grupos ou categorias distintas. Este trabalho explora quatro variáveis categóricas: solvente, ordem de extração, tempo e Vm-ratio. Aplicamos a codificação binária para representá-las de forma adequada em análises estatísticas.

Variáveis Categóricas do problema:

- Solvente: Possui 3 níveis: 1, 2 e 3.
- Ordem de Extração: Tem 3 níveis: 1, 2 e 3.
- Tempo: Apresenta 2 níveis: 15 e 60.
- Vm-ratio: Contém 3 níveis: 5, 10 e 20.

Codificação Binária (One-Hot Encoding)

A codificação binária é uma técnica utilizada para representar variáveis categóricas, onde cada categoria é representada por uma coluna binária. O valor 1 indica a presença da categoria e o valor 0 indica a ausência.

Podemos verificar abaixo uma parte do DataFrame já com as variáveis categóricas:

Index	Order_1,0	Order_2,0	Order_3,0	Time_15,0	Time_60,0	vm-ratio_5,0	vm-ratio_10,0	vm-ratio_20,0	Solvent_1,0	Solvent_2,0	Solvent_3,0
0	1	0	0	1	0	1	0	0	1	0	0
1	1	0	0	1	0	1	0	0	1	0	0
2	1	0	0	1	0	1	0	0	1	0	0
3	1	0	0	1	0	1	0	0	1	0	0
4	1	0	0	1	0	1	0	0	1	0	0
5	1	0	0	1	0	1	0	0	1	0	0
6	0	1	0	1	0	1	0	0	1	0	0
7	0	1	0	1	0	1	0	0	1	0	0
8	0	1	0	1	0	1	0	0	1	0	0
9	0	1	0	1	0	1	0	0	1	0	0
10	0	1	0	1	0	1	0	0	1	0	0
11	0	0	1	1	0	1	0	0	1	0	0
12	0	0	1	1	0	1	0	0	1	0	0
13	0	0	1	1	0	1	0	0	1	0	0
14	0	0	1	1	0	1	0	0	1	0	0
15	0	0	1	1	0	1	0	0	1	0	0
16	1	0	0	1	0	0	1	0	1	0	0
17	1	0	0	1	0	0	1	0	1	0	0
18	1	0	0	1	0	0	1	0	1	0	0
19	1	0	0	1	0	0	1	0	1	0	0
20	1	0	0	1	0	0	1	0	1	0	0
21	1	0	0	1	0	0	1	0	1	0	0
22	0	1	0	1	0	0	1	0	1	0	0
23	0	1	0	1	0	0	1	0	1	0	0
24	0	1	0	1	0	0	1	0	1	0	0
25	0	1	0	1	0	0	1	0	1	0	0
26	0	1	0	1	0	0	1	0	1	0	0
27	0	0	1	1	0	0	1	0	1	0	0
28	0	0	1	1	0	0	1	0	1	0	0
29	0	0	1	1	0	0	1	0	1	0	0

Figura 14: Gráfico de Dispersão (Var.Saída vs Var.Entrada)

Para aplicar a codificação binária, vamos seguir estes passos para cada variável:

Solvente

- Metanol: (1, 0, 0)
- Etanol: (0, 1, 0)
- Acetona: (0, 0, 1)

Ordem de Extração

- 1ª extração: (1, 0, 0)
- 2ª extração: (0, 1, 0)
- 3ª extração: (0, 0, 1)

Tempo

- 15 minutos: (1, 0)
- 60 minutos: (0, 1)

Vm-ratio

- 25 mL/5g: (1, 0, 0)
- 50 mL/5g: (0, 1, 0)
- 100 mL/5g: (0, 0, 1)

Conclusão

A codificação binária é uma ferramenta útil para lidar com variáveis categóricas, permitindo uma representação clara e distinta dos dados. Ao aplicar a codificação binária às variáveis solvente, ordem de extração, tempo e Vm-ratio, podemos preparar os dados para análises estatísticas e modelagem de machine learning de forma adequada.

2.4.6 Normalização:

A normalização é uma etapa crucial no processo de mineração de dados descrito pelo CRISP-DM (Cross Industry Standard Process for Data Mining). O objetivo da normalização é transformar dados em um formato que possa ser eficientemente processado por algoritmos de aprendizado de máquina, melhorando a qualidade dos resultados.

Objetivos da Normalização

A normalização de dados é crucial para otimizar o desempenho dos algoritmos de aprendizado de máquina, abordando a redução de variações de escala, o aprimoramento da convergência e a prevenção de sobrecarga numérica.

Tipos de Normalização

Existem várias técnicas de normalização, cada uma adequada a diferentes tipos de dados e objetivos de análise. Aqui estão algumas das mais comuns:

1. O MinMaxScaler é uma técnica de normalização usada para dimensionar os dados num intervalo específico, geralmente entre 0 e 1. O objetivo é garantir que todas as características tenham a mesma escala, o que pode ser útil para algoritmos de aprendizagem automática que são sensíveis à escala dos dados.

2. A normalização Z-score, ou padronização, é uma técnica comum de normalização usada para ajustar os dados de forma a terem média zero e desvio padrão um. Aplica-se a cada característica dos dados, resultando numa média de zero e um desvio padrão de um para cada característica.
3. `StandardScaler()`: Padroniza os recursos removendo a média e escalonando para a variância unitária.

Relevância do `StandardScaler()`

O `StandardScaler()` é uma técnica comum de normalização utilizada em aprendizagem automática. É particularmente útil quando as características têm distribuições gaussianas e quando a escala das características não é conhecida antecipadamente.

Abaixo está todo o código usado bem para fazer a normalização:

```
# In[] Normalização

# Especificar as variáveis de entrada (X) e saída (y)
X_cols = ['Solvent', 'Order', 'Time', 'Vm-ratio'] # Defina as variáveis de entrada
y_cols = ['TPC', 'AOA', 'ANT'] # Defina as variáveis de saída

# Separar as variáveis de entrada e saída
X = df[X_cols]
y = df[y_cols]

# Normalizar os dados de entrada
scaler = StandardScaler()
X_normalized = scaler.fit_transform(X)

# Divisão dos dados em conjunto de treino e teste
X_train, X_test, y_train, y_test = train_test_split(X_normalized, y, test_size=0.2, random_state=42)
```

Figura 15: Normalização:

2.5 Modelação dos dados

Na fase de modelação de dados, diversos algoritmos são aplicados aos dados preparados para criar modelos preditivos. A escolha do algoritmo adequado é vital para obter insights precisos e relevantes. Esta etapa envolve a aplicação de diferentes técnicas de aprendizagem automática, ajustando os seus parâmetros e avaliando as suas performances com base em métricas específicas. A modelação de dados é uma etapa crucial no processo de análise preditiva, onde diferentes modelos estatísticos e de aprendizagem automática são treinados para prever variáveis de interesse com base num conjunto de dados.

Este trabalho descreve o processo de modelação de dados usando três abordagens distintas: Random Forest Regressor, Principal Component Analysis (PCA) com Random Forest Regressor, e Regressão Linear. Cada modelo foi avaliado com base em várias métricas de desempenho, e os resultados foram comparados para determinar a melhor abordagem para o conjunto de dados em questão.

Modelos Utilizados

A seguir serão apresentados e explicados os modelos utilizados:

- RandomForestRegressor:

O RandomForestRegressor é um método de aprendizado ensemble que utiliza múltiplas árvores de decisão para melhorar a precisão preditiva e controlar o overfitting. Este algoritmo é robusto e eficaz para uma variedade de problemas de regressão. Suas vantagens incluem a capacidade de lidar com dados de alta dimensionalidade e uma variedade de tipos de variáveis, além de ser menos sensível a outliers em comparação com outros modelos. No entanto, pode ser computacionalmente intensivo para conjuntos de dados muito grandes.

- PCA com RandomForestRegressor:

A Análise de Componentes Principais (PCA) é uma técnica de redução de dimensionalidade que transforma os dados em um novo conjunto de variáveis não correlacionadas, chamadas de componentes principais. Aplicar PCA antes de um modelo de RandomForestRegressor pode melhorar a eficiência computacional e a performance do modelo ao eliminar colinearidades. Isso pode ser especialmente útil quando lidamos com conjuntos de dados com muitas variáveis correlacionadas, ajudando a simplificar o modelo e reduzir o overfitting.

- Regressão Linear:

A Regressão Linear é um modelo estatístico simples que assume uma relação linear entre a variável dependente e uma ou mais variáveis independentes. É um método interpretável e fácil de implementar, sendo útil para entender a relação direta entre as variáveis. No entanto, pode ser limitado em termos de capacidade preditiva para dados complexos, especialmente quando a relação entre as variáveis não é estritamente linear. Também pode ser sensível a outliers e violações das suposições do modelo, como a normalidade dos resíduos.

Métricas

As métricas de avaliação são fundamentais para medir o desempenho dos modelos estatísticos em relação aos dados observados. No campo da ciência de dados e aprendizado de máquina, estas métricas, como o Erro Quadrático Médio (MSE) e o Erro Absoluto Médio (MAE), são essenciais para avaliar a precisão das previsões.

Cada métrica oferece uma perspectiva única sobre o desempenho do modelo. Por exemplo, o MSE mede a média dos quadrados dos erros, enquanto o MAE mede a média dos valores absolutos dos erros. Compreender essas métricas ajuda os profissionais a escolher e otimizar os modelos para suas necessidades específicas.

Erro Quadrático Médio (MSE - Mean Squared Error):

O MSE é a média dos quadrados dos erros entre os valores previstos pelo modelo e os valores reais do conjunto de teste.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Onde:

- n é o número de observações no conjunto de teste
- y_i é o valor real da i -ésima observação,
- \hat{y}_i é o valor previsto para a i -ésima observação.

Raiz do Erro Quadrático Médio (RMSE - Root Mean Squared Error):

O RMSE é a raiz quadrada do MSE e fornece uma interpretação da média dos erros em unidades da variável de destino.

$$RMSE = \sqrt{MSE}$$

Coefficiente de Determinação (R^2 - R-squared):

O R^2 é uma medida estatística que representa a proporção da variância na variável dependente que é previsível a partir das variáveis independentes no modelo. Ele fornece uma indicação do ajuste do modelo aos dados observados e varia entre 0 e 1.

$$R^2 = 1 - \frac{SSE}{SST}$$

Onde:

- SSE é a soma dos quadrados dos resíduos (erros),
- SST é a soma total dos quadrados.

Erro Absoluto Médio (MAE - Mean Absolute Error):

O MAE é a média dos valores absolutos dos erros entre os valores previstos e os valores reais.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Erro Percentual Absoluto Médio (MAPE - Mean Absolute Percentage Error):

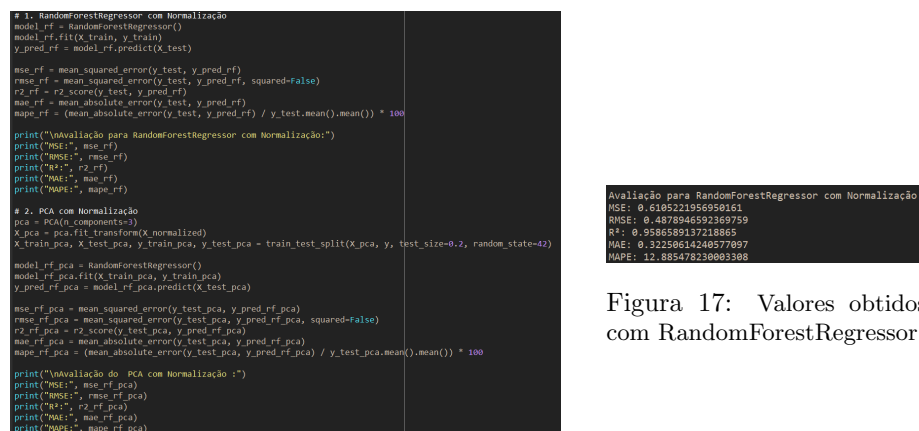
O MAPE é uma métrica expressa em porcentagem, que calcula a média das razões entre os erros absolutos e os valores reais.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Estas métricas são frequentemente usadas para avaliar o desempenho de modelos de regressão. O objetivo é minimizar essas métricas para obter previsões mais precisas e confiáveis.

Após uma explicação detalhada sobre os modelos escolhidos e as métricas que serão usadas para este estudo, segue-se uma explicação em nível de Python de todo o código, bem como dos resultados obtidos.

Para começar, foi selecionado o modelo RandomForest com dados normalizados. A partir deste modelo, obtivemos os seguintes valores:



```
# 1. RandomForestRegressor com Normalização
model_rf = RandomForestRegressor()
model_rf.fit(X_train, y_train)
y_pred_rf = model_rf.predict(X_test)

mse_rf = mean_squared_error(y_test, y_pred_rf)
rmse_rf = mean_squared_error(y_test, y_pred_rf, squared=False)
r2_rf = r2_score(y_test, y_pred_rf)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
mape_rf = (mean_absolute_error(y_test, y_pred_rf) / y_test.mean().mean()) * 100

print("Avaliação para RandomForestRegressor com Normalização:")
print("MSE:", mse_rf)
print("RMSE:", rmse_rf)
print("R2:", r2_rf)
print("MAE:", mae_rf)
print("MAPE:", mape_rf)

# 2. PCA com Normalização
pca = PCA(n_components=3)
X_pca = pca.fit_transform(X_normalizad)
X_train_pca, X_test_pca, y_train_pca, y_test_pca = train_test_split(X_pca, y, test_size=0.2, random_state=42)

model_rf_pca = RandomForestRegressor()
model_rf_pca.fit(X_train_pca, y_train_pca)
y_pred_rf_pca = model_rf_pca.predict(X_test_pca)

mse_rf_pca = mean_squared_error(y_test_pca, y_pred_rf_pca)
rmse_rf_pca = mean_squared_error(y_test_pca, y_pred_rf_pca, squared=False)
r2_rf_pca = r2_score(y_test_pca, y_pred_rf_pca)
mae_rf_pca = mean_absolute_error(y_test_pca, y_pred_rf_pca)
mape_rf_pca = (mean_absolute_error(y_test_pca, y_pred_rf_pca) / y_test_pca.mean().mean()) * 100

print("Avaliação do PCA com Normalização:")
print("MSE:", mse_rf_pca)
print("RMSE:", rmse_rf_pca)
print("R2:", r2_rf_pca)
print("MAE:", mae_rf_pca)
print("MAPE:", mape_rf_pca)
```

```
Avaliação para RandomForestRegressor com Normalização:
MSE: 0.6105221956950161
RMSE: 0.4878946592369759
R2: 0.9586589137218005
MAE: 0.32250614240577097
MAPE: 12.885478230003308
```

Figura 17: Valores obtidos com RandomForestRegressor

Figura 16: Código RandomForestRegressor com Normalização

Vamos analisar os resultados de cada modelo em relação ao problema em questão:

RandomForestRegressor com Normalização:

- MSE (Erro Quadrático Médio): 0,6105
- RMSE (Raiz do Erro Quadrático Médio): 0,4879
- R² (Coeficiente de Determinação): 0,9587

- MAE (Erro Médio Absoluto): 0,3225
- MAPE (Erro Percentual Absoluto Médio): 12,8855

Os resultados do modelo RandomForestRegressor com normalização revelam um desempenho excepcional. Com valores baixos de MSE e RMSE, indica-se uma precisão elevada nas previsões, enquanto o alto valor de R^2 sugere um ajuste quase perfeito aos dados. Além disso, os valores de MAE e MAPE indicam erros absolutos e percentuais baixos, respectivamente, reforçando a confiabilidade e utilidade prática do modelo. Em suma, o modelo demonstra uma capacidade notável de previsão e é altamente eficaz para as tarefas propostas.

PCA com Normalização:

- MSE: 0,5974
- RMSE: 0,4798
- R^2 : 0,9617
- MAE: 0,3373
- MAPE: 13,4779

Os valores de MSE e RMSE do modelo PCA com normalização indicam um erro baixo nas previsões, sugerindo uma boa precisão. Com um R^2 de 0,9617, o modelo explica quase toda a variância dos dados, demonstrando um excelente ajuste. Além disso, os valores de MAE e MAPE são baixos, o que significa que as previsões são precisas tanto em termos absolutos quanto relativos.

Em resumo, o modelo PCA com normalização mostra um desempenho excelente, com altos níveis de precisão e um ótimo ajuste aos dados, sendo uma escolha robusta para previsões confiáveis.

Regressão Linear com Normalização:

- MSE: 0,0266
- RMSE: 0,1631
- R^2 : 0,4079
- MAE: 0,1372
- MAPE: 58,9841

Os valores de MSE e RMSE do modelo de Regressão Linear com normalização indicam um erro relativamente baixo nas previsões. Entretanto, o valor de R^2 , que é 0,4079, mostra que o modelo explica menos da metade da variância dos dados, sugerindo um ajuste modesto. Apesar do MAE ser baixo, o MAPE é bastante elevado, indicando que, apesar dos erros absolutos serem pequenos, os erros percentuais em relação aos valores reais são altos. Isso sugere que o modelo pode não ser tão preciso em termos relativos.

Em resumo, o modelo de Regressão Linear com normalização apresenta um desempenho misto. Apesar dos erros absolutos serem baixos, o ajuste aos dados é modesto e os erros percentuais são elevados, sugerindo que o modelo pode não ser a melhor escolha para previsões precisas em termos relativos.

Conclusão:

Este estudo compara o desempenho de diferentes modelos de regressão usando diversas métricas de avaliação, como MSE, RMSE, R^2 , MAE e MAPE. Os resultados indicam que o modelo RandomForestRegressor com normalização supera os outros em precisão e ajuste aos dados, seguido pelo modelo PCA com normalização. No entanto, a Regressão Linear com normalização mostra um desempenho inferior. A seleção do modelo ideal depende da natureza dos dados, das necessidades de precisão e dos recursos computacionais disponíveis. Este estudo ressalta a importância de experimentar e explorar diferentes modelos e técnicas de avaliação para desenvolver soluções eficazes em problemas de regressão em ciência de dados.

2.6 Avaliação:

A etapa de Avaliação no CRISP-DM é crucial para verificar a qualidade e a eficácia dos modelos criados na fase de modelagem. Mesmo que um modelo apresente bom desempenho em termos de métricas estatísticas, precisa ser validado para garantir previsões precisas, úteis e alinhadas com os objetivos do negócio. Nesta etapa, os analistas de dados avaliam o modelo em termos de precisão, robustez e generalização, usando conjuntos de dados de teste não utilizados durante o treino.

Os principais objetivos desta fase são avaliar o desempenho do modelo utilizando métricas como MSE (Erro Quadrático Médio), RMSE (Raiz do Erro Quadrático Médio), R^2 (Coeficiente de Determinação), MAE (Erro Absoluto Médio) e MAPE (Erro Absoluto Percentual Médio). É essencial assegurar que o modelo cumpre os requisitos do negócio e identificar áreas de melhoria, seja através do ajuste de hiperparâmetros, da seleção de variáveis ou da escolha de diferentes algoritmos. Além disso, é importante validar a generalização do modelo, garantindo que funcione bem com dados não vistos.

Durante a Avaliação, diferentes modelos e versões são comparados utilizando estas métricas para determinar qual oferece o melhor desempenho global, ajudando na escolha do modelo mais adequado para os objetivos do projeto.

Comparação das Métricas de Avaliação dos Modelos:



Figura 18: Normalização:

O gráfico acima apresenta uma comparação das métricas de avaliação para três diferentes métodos de modelagem: Normalização, Análise de Componentes Principais (PCA) e Regressão Linear. As métricas avaliadas incluem MSE, RMSE, R^2 , MAE e MAPE, permitindo uma análise detalhada do desempenho de cada modelo. Esta visualização ajuda a identificar o modelo que melhor atende aos critérios de precisão e robustez, essenciais para uma implementação eficaz.

Conclusões dos gráficos obtidos:

MSE (Erro Quadrático Médio):

- Normalização: 0.56
- PCA: 0.61
- Regressão Linear: 0.03

A Regressão Linear apresentou o menor MSE, indicando melhor desempenho neste critério específico.

RMSE (Raiz do Erro Quadrático Médio):

- Normalização: 0.47
- PCA: 0.49
- Regressão Linear: 0.16

Similar ao MSE, a Regressão Linear também tem o menor RMSE, corroborando seu desempenho superior.

R² (Coeficiente de Determinação):

- Normalização: 0.96
- PCA: 0.96
- Regressão Linear: 0.41

Tanto a Normalização quanto a PCA têm valores R² muito altos (0.96), indicando uma alta proporção de variância explicada pelo modelo. A Regressão Linear tem um R² significativamente menor (0.41).

MAE (Erro Absoluto Médio):

- Normalização: 0.31
- PCA: 0.33
- Regressão Linear: 0.14

A Regressão Linear novamente apresenta o menor valor de MAE, mostrando menor erro absoluto médio.

MAPE (Erro Absoluto Percentual Médio):

- Normalização: 12.44
- PCA: 13.1
- Regressão Linear: 58.98

Aqui, a Normalização e a PCA têm valores de MAPE relativamente próximos e baixos, enquanto a Regressão Linear apresenta um valor extremamente alto (58.98), indicando que, em termos percentuais, os erros são muito maiores para a Regressão Linear.

Conclusões:

A perspectiva a que se pode chegar é que a Regressão Linear não seria um bom modelo devido ao facto de o MAE (Erro Médio Absoluto) possuir um valor demasiado alto e o R^2 um valor demasiado baixo. Assim, temos de comparar os modelos restantes: o RandomForestRegressor e o PCA.

Podemos concluir que o modelo Random Forest geralmente apresenta melhores resultados em termos de MSE (Erro Quadrático Médio), RMSE (Raiz do Erro Quadrático Médio), MAE (Erro Médio Absoluto) e MAPE (Erro Percentual Absoluto Médio), indicando uma melhor precisão preditiva em relação aos erros quadráticos médios, suas raízes, erros absolutos médios e erros percentuais médios. Embora o PCA tenha uma leve vantagem no R^2 , essa técnica é mais comumente utilizada para redução de dimensionalidade, não sendo propriamente um modelo preditivo.

Assim, considerando as métricas fornecidas, o Random Forest seria o melhor modelo para o estudo, oferecendo um desempenho superior em termos de erros preditivos e precisão geral.

3 Resultados e Análise

A análise detalhada realizada neste estudo sobre as amostras de beterraba destacou a importância das técnicas estatísticas e de modelagem de dados na compreensão da composição e das propriedades dos compostos bioativos presentes nesta planta. Exploramos a relevância desses compostos para a saúde e a necessidade crucial de extrair essas substâncias de forma eficiente.

Por meio de uma variedade de métodos estatísticos e de visualização, investigamos as relações entre diferentes variáveis, como solvente, razão volume-massa, ordem e tempo de extração, e as propriedades dos extratos, como compostos fenólicos totais, antocianinas e atividade antioxidante. Essa análise revelou padrões, tendências e correlações significativas, fornecendo insights valiosos para orientar o processo de extração e melhorar a qualidade dos extratos obtidos.

A aplicação de técnicas de modelagem de dados, como o RandomForestRegressor e o PCA, possibilitou o desenvolvimento de modelos preditivos robustos para prever variáveis importantes relacionadas às propriedades das amostras de beterraba. A avaliação desses modelos usando métricas como MSE, RMSE, R^2 , MAE e MAPE identificou o modelo RandomForestRegressor com normalização como a melhor opção, fornecendo previsões precisas e confiáveis. Além disso, a fase de implementação ressaltou a importância de garantir uma transição suave das soluções desenvolvidas para o ambiente de produção. Isso permitirá que os resultados obtidos durante a análise estatística e modelagem de dados sejam aplicados de forma eficaz pelos utilizadores finais, com potenciais aplicações na indústria alimentícia para melhorar a produção de alimentos funcionais ou suplementos nutricionais. No entanto, para ampliar ainda mais o impacto deste estudo, seria importante considerar futuras direções de pesquisa. Investigações adicionais sobre os efeitos dos compostos bioativos da beterraba na saúde humana e a otimização contínua de técnicas de extração poderiam proporcionar avanços significativos neste campo.

Por fim, este estudo integra eficazmente a análise detalhada das propriedades das amostras de beterraba com o uso de métodos estatísticos e de modelagem de dados, fornecendo insights valiosos e desenvolvendo soluções preditivas úteis. Ao associar a informação específica das amostras de beterraba com o método e as métricas utilizadas, este trabalho contribui significativamente para o avanço do conhecimento no campo da extração de compostos bioativos de plantas e demonstra o potencial das técnicas analíticas para apoiar a pesquisa e a inovação em nutrição e saúde.

Referências

- [1] Cross Industry Standard Process for Data Mining. *Wikipedia*. Disponível em: <https://pt.wikipedia.org/wiki/CrossIndustryStandardProcessforDataMining>. Acesso em : 21 de maio de 2024.
- [2] Random Forest. *IBM*. Disponível em: <https://www.ibm.com/topics/random-forest><https://www.ibm.com/topics/random-forest>. Acesso em: 21 de maio de 2024.
- [3] Linear Regression. *IBM*. Disponível em: <https://www.ibm.com/topics/linear-regression><https://www.ibm.com/topics/linear-regression>. Acesso em: 21 de maio de 2024.
- [4] O que é uma Matriz de Correlação?. *Data Science*. Disponível em: <https://datascience.eu/pt/matematica-e-estatistica/o-que-e-uma-matriz-de-correlacao/>. Acesso em: 21 de maio de 2024.
- [5] Usamos todos os recursos fornecidos pelo docente para a realização deste trabalho.