1 Introdução

O uso do solo de maneira adequada é de fundamental importância para maximizar o retorno do plantio e manter o solo em boas condições para que possa ser utilizado por vários anos sem degradação. Caso o solo seja mal manejado, pode-se acabar com um terreno infértil, o que aumenta a demanda de recursos para o cultivo e recuperação da área de plantio.

Conforme o solo é utilizado para o plantio de um tipo de alimento, ocorre a diminuição do tipo de nutrientes consumidos por essa planta, e em contrapartida os outros se tornam abundantes pelo acúmulo durante o tempo em que não foi consumido. Dessa forma, pode-se realizar a rotação de culturas, plantando alimentos que consomem nutrientes diferentes, assim o solo se mantém mais bem preservado.

Com objetivo de facilitar a escolha da cultura a ser semeada em um terreno específico será feito o treinamento de um modelo de aprendizado supervisionado, para isso, serão utilizados dados que consistem em características do terreno, principalmente relacionadas à quantidade de nutrientes e substâncias nele presentes, e a classificação será feita de acordo com um tipo de alimento que é considerado ideal para o solo em questão.

A tabela de dados apresenta 2200 diferentes terrenos e suas respectivas culturas ideais, as quais são divididas em 22 classes que indicam diferentes sementes e frutas, para cada um desses terrenos temos valores de quantidade de nitrogênio, fósforo, potássio, além de temperatura, umidade e pH, e finalmente temos uma estimativa da quantia de chuva que a plantação receberá durante o crescimento.

2 Metodologia

2.1 Origem dos Dados

Os dados foram obtidos diretamente da plataforma *kaggle*, um site para estudo de ciência de dados e machine learning, e podem ser obtidos através do link https://www.kaggle.com/datasets/aksahaha/crop-recommendation. Segundo o usuário Abhishek Kumar, que disponibilizou os dados, eles são provenientes do ICAR (Indian Council of Agriculture Research), e complementados por pesquisas na internet feitas por ele.

2.2 Dicionário dos Dados

- **Nitrogênio (*nitrogen)*:** Representa a quantidade de nitrogênio (em kg/ha) presente no solo para a cultura. O nitrogênio é um nutriente essencial para o crescimento de plantas, e sua deficiência ou excesso pode afetar o crescimento e a produção da cultura;
- **Fósforo (*phosphorus*):** Representa a quantidade de fósforo (em kg/ha) presente no solo para a cultura. Também é um elemento essencial no plantio, sendo importante para processos como transferência de energia e fotossíntese;
- **Potássio (*potassium)*:** Representa a quantidade de potássio (em kg/ha) presente no solo para a cultura. Também é um elemento essencial, e é importante para processos fisiológicos como regulação de água e transporte de nutrientes;
- **Temperatura (*temperature)*:** Representa a temperatura média (em Celsius) durante o período de crescimento da cultura. A temperatura é um fator ambiental importante que pode afetar o crescimento e o desenvolvimento das plantas, e cada cultura possui uma temperatura ideal;
- **Humidade (*humidity)*:** Representa a humidade relativa (em porcentagem) durante o período de crescimento da cultura. A humidade é outro fator ambiental importante, tendo em vista que uma alta humidade pode promover a proliferação de fungos e desenvolvimento de doenças;
- **pH:** Representa o pH da cultura durante seu período de crescimento. O pH é uma medida de acidez ou alcalinidade do solo e pode afetar a disponibilidade de nutrientes para a cultura;

- **Precipitação (*rainfall*):** Representa a precipitação (em mm) durante o período de crescimento da cultura. Cada cultura necessita de uma quantidade diferente de água, o que torna a precipitação outro fator ambiental importante;
- **Rótulo (*label*):** Representa o tipo da cultura.

3 Análise dos Dados

3.1 Medidas Descritivas

Tabela 1: Medidas Descritivas

	Nitrogênio	Fósforo	Potássio	Temperatura	Humidade	pH	Precipitação
Mínimo	0.00	5.00	5.00	8.82	14.25	3.50	20.21
Máximo	140.00	145.00	205.00	43.675	99.98	9.93	298.56
Média	50.55	53.36	48.14	25.61	71.48	6.46	103.46
Mediana	37.00	51.00	32.00	25.59	80.47	6.42	94.86
Variância	1362.89	1088.06	2565.21	25.64	495.67	0.59	3020.42
Desvio-padrão	36.91	32.98	50.64	5.06	22.26	0.77	54.95
Coeficiente de Variância	73.02	61.81	105.19	19.76	31.14	11.96	53.11
Amplitude	140.00	140.00	200.00	34.85	85.72	6.43	278.34

Fonte: Autoria própria

A partir da tabela 1, pode-se ter uma ideia inicial das distribuições das características. É possível inferir que as variáveis *temperature* e pH possuem uma curva simétrica, já que suas médias e medianas são bem próximas, enquanto *humidity* provavelmente possui uma curva assimétrica à esquerda. Para todos os outros atributos as curvas são possivelmente assimétricas à direita.

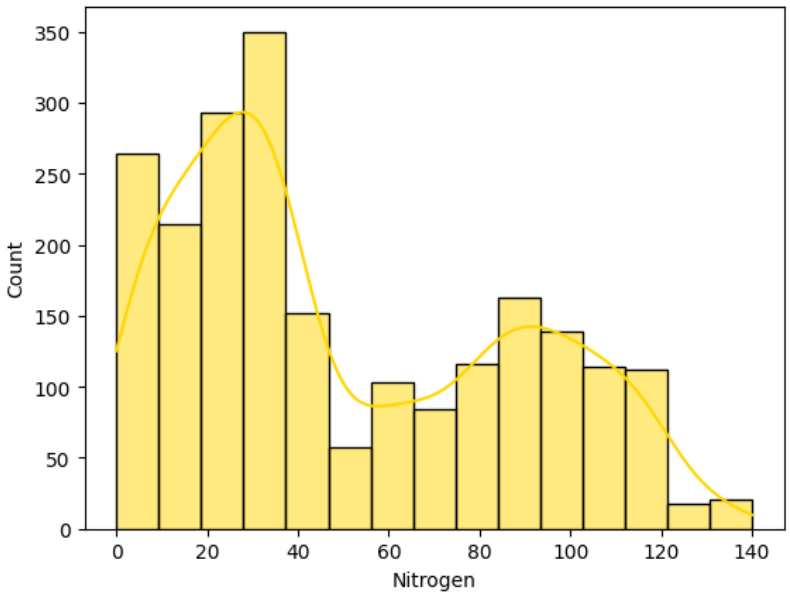
No que tange as medidas de dispersão, a análise anterior é reforçada. As medidas de variância e desvio-padrão apresentam valores altos para as variáveis que não são simétricas, o que indica que há uma alta variabilidade nos dados. Ou seja, há valores que possuem uma grande distância da média.

3.2 Visualização dos Dados

3.2.1 Histogramas

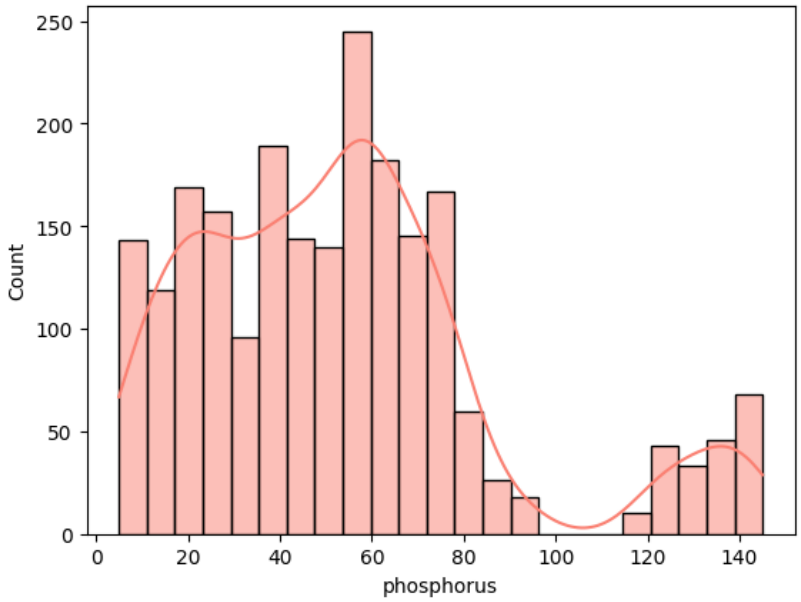
A partir da figura 1, é possível observar o formato da curva dos atributos. Assim, há mais evidências de que a análise anteriormente feita está, provavelmente, correta. Fazem-se necessários, então, testes de hipóteses.

Figura 1: Distribuição da varíavel *nitrogen*



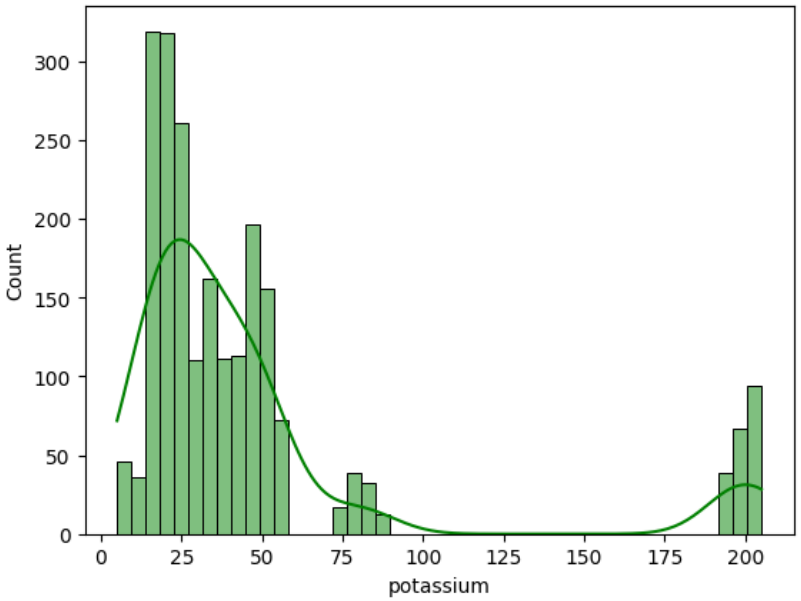
Fonte: Autoria própria

Figura 2: Distribuição da variável *phosphorus*



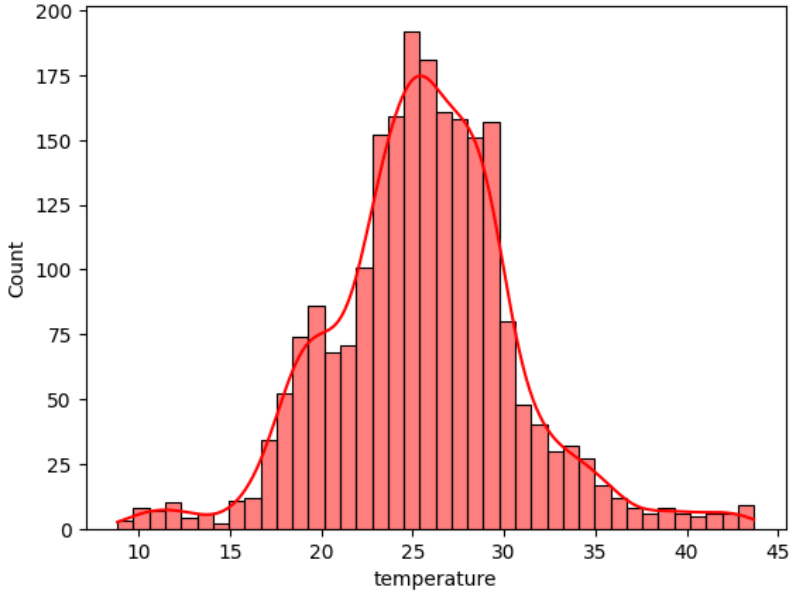
Fonte: Autoria própria

Figura 3: Distribuição da variável *potassium*



Fonte: Autoria própria

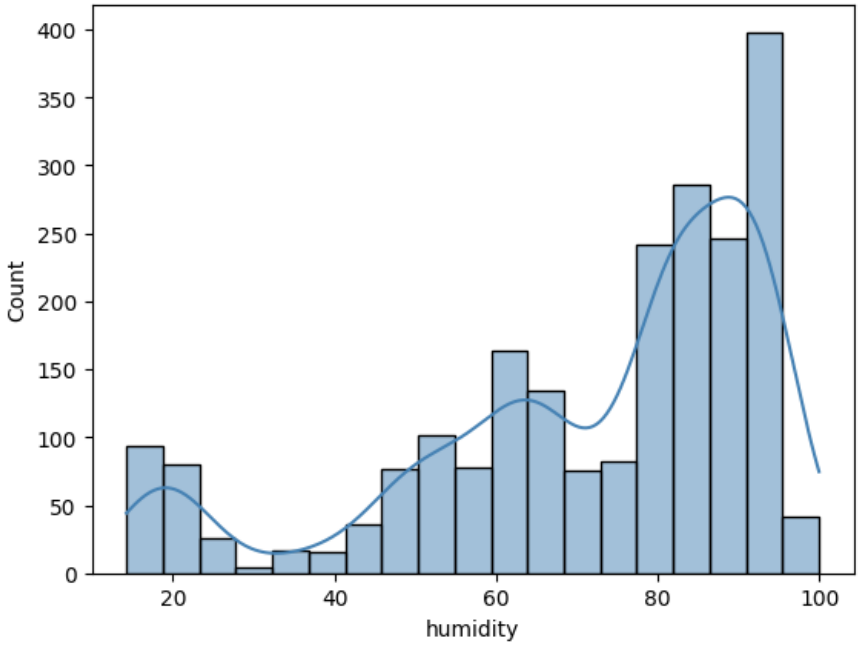
Figura 4: Distribuição da variável *temperature*



Fonte: Autoria própria

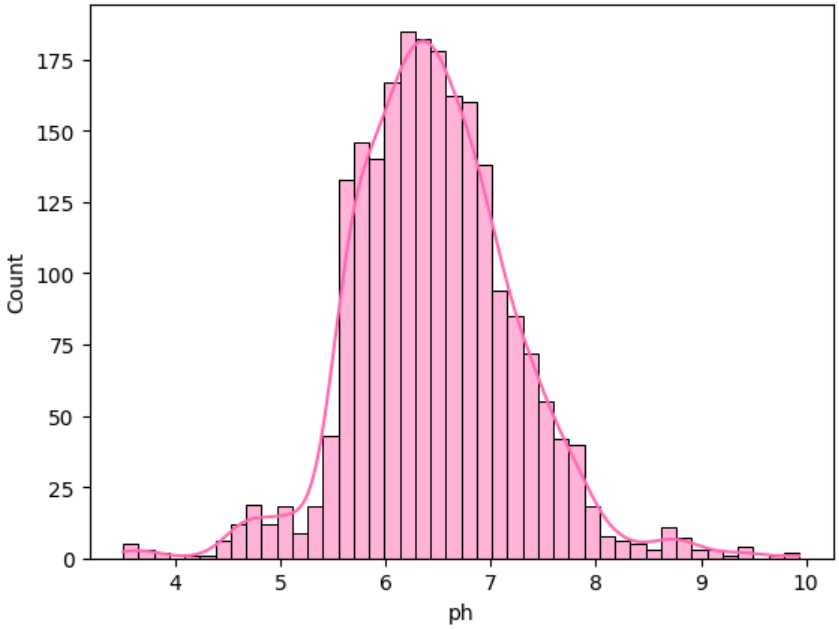
Percebe-se, como anteriormente dito, que a distribuição dos dados referentes à variável temperatura provavelmente segue uma distribuição.

Figura 5: Distribuição da variável *humidity*



Fonte: Autoria própria

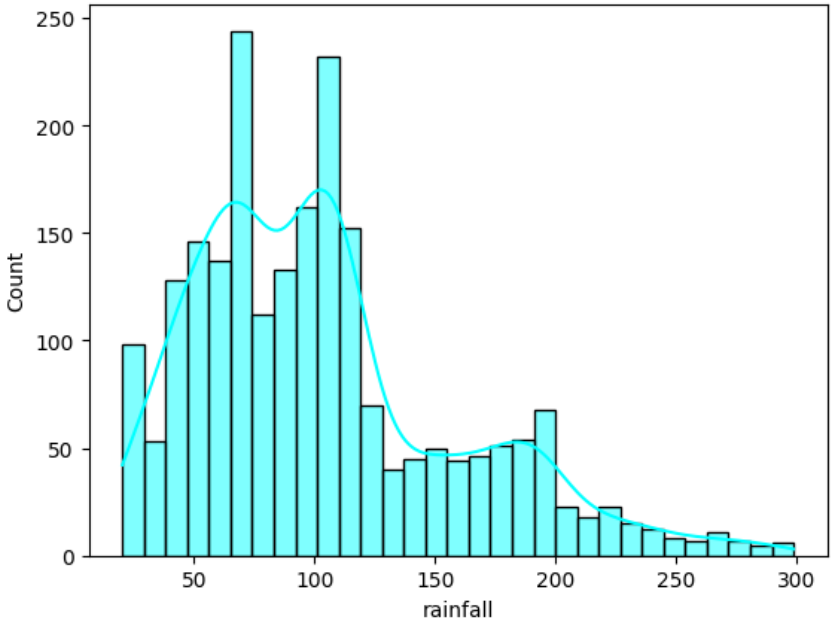
Figura 6: Distribuição da variável *pH*



Fonte: Autoria própria

Percebe-se, como anteriormente dito, que a distribuição dos dados referentes à variável pH provavelmente segue uma distribuição.

Figura 7: Distribuição da variável *rainfall*

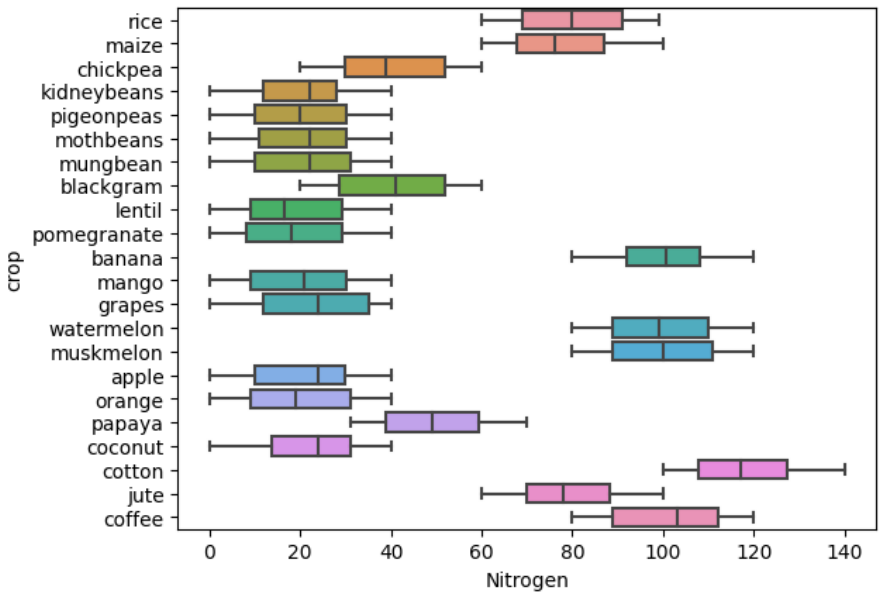


Fonte: Autoria própria

3.2.2 Boxplots

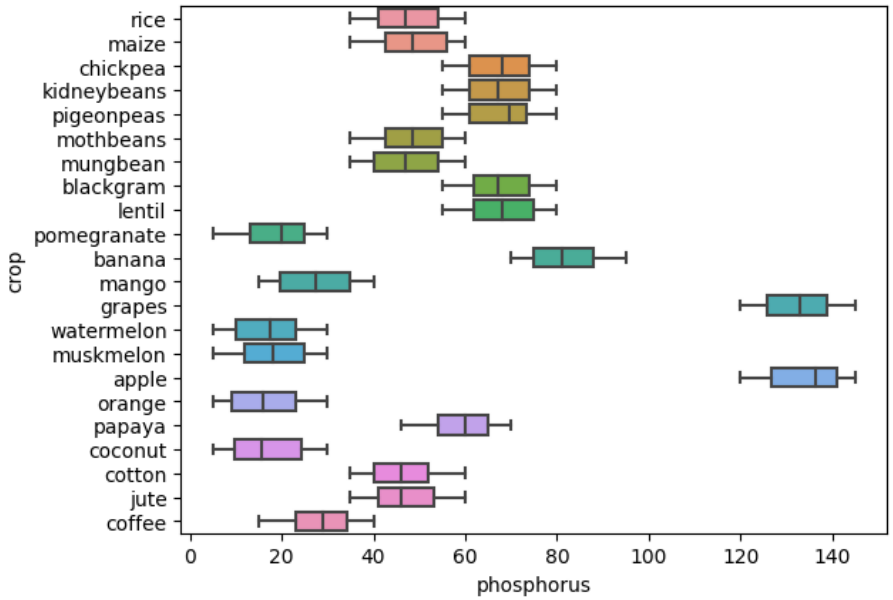
Com os *boxplots*, é possível comparar a distribuição dos dados em relação ao atributo-alvo. Mais uma vez, a teoria de que as variáveis *pH* e *temperature* são mais balanceadas é corroborada.

Figura 8: Relação entre o rótulo e o nitrogênio



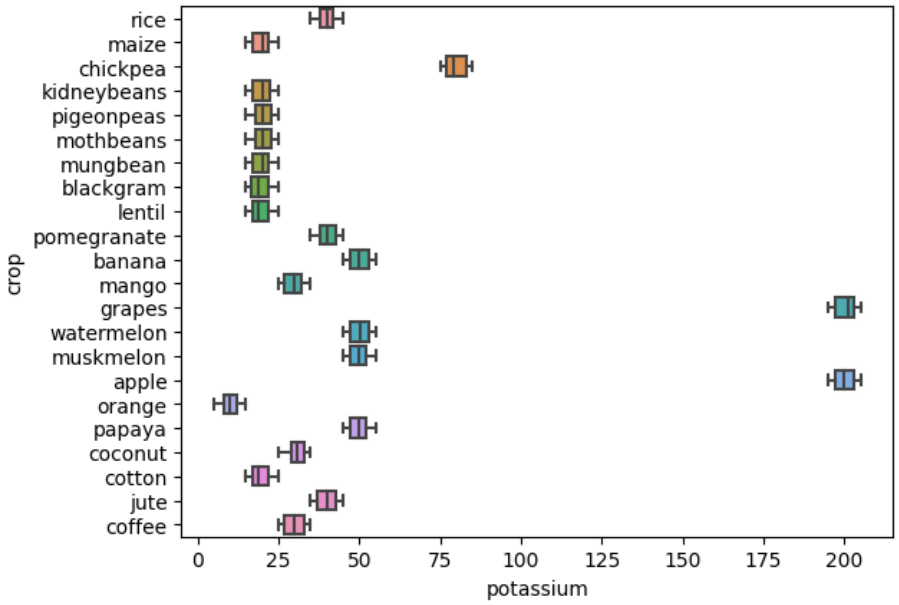
Fonte: Autoria própria

Figura 9: Relação entre o rótulo e fósforo



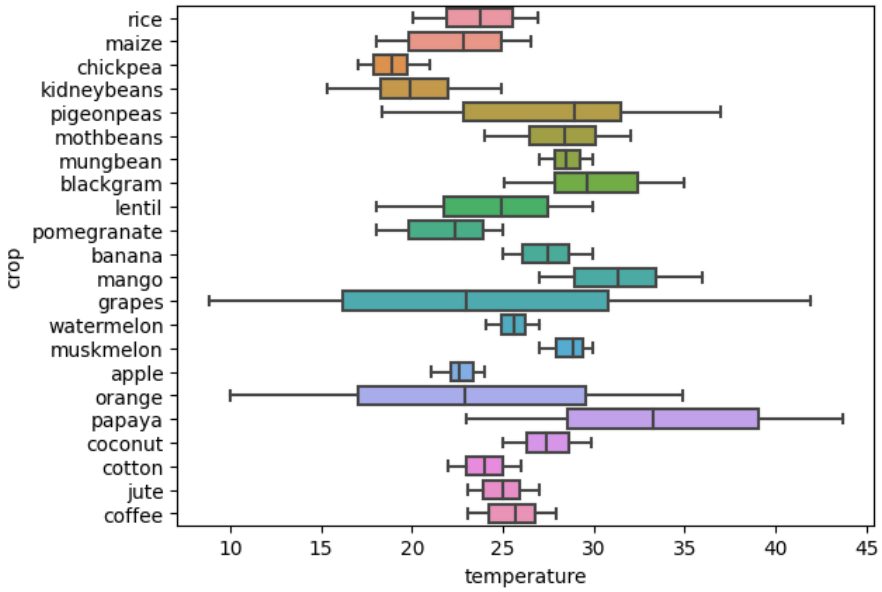
Fonte: Autoria própria

Figura 10: Relação entre rótulo e potássio



Fonte: Autoria própria

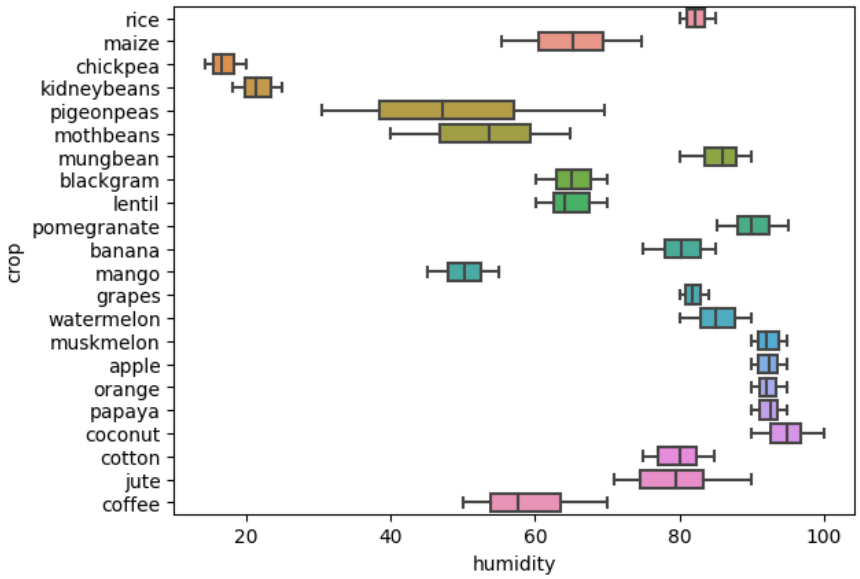
Figura 11: Relação entre rótulo e temperatura



Fonte: Autoria própria

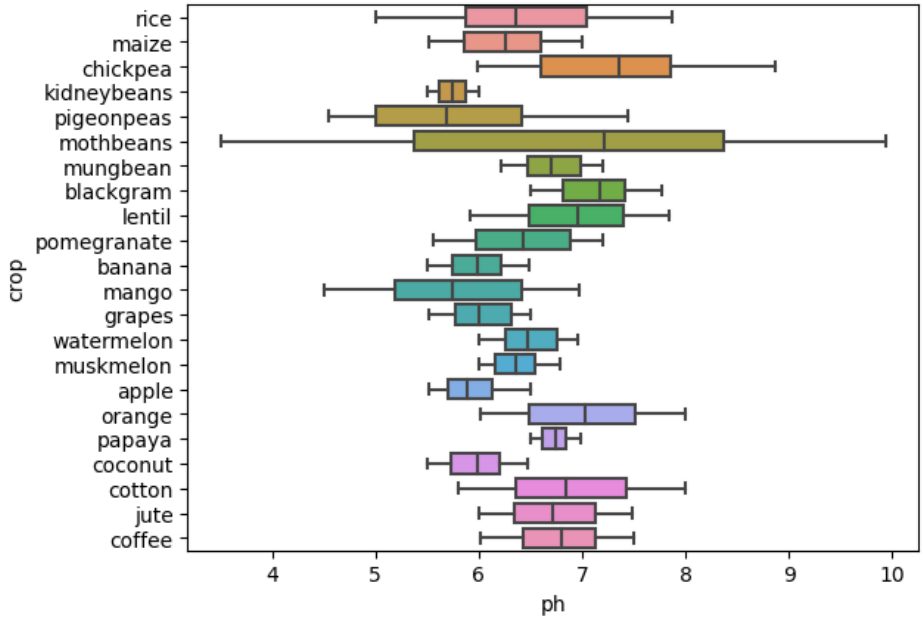
A maior parte dos valores está ao redor da média, que é de aproximadamente 25. Apesar de certos valores apresentarem grande variação, como *grapes* e *orange*, isso não afetou a curva.

Figura 12: Relação entre rótulo e humidade



Fonte: Autoria própria

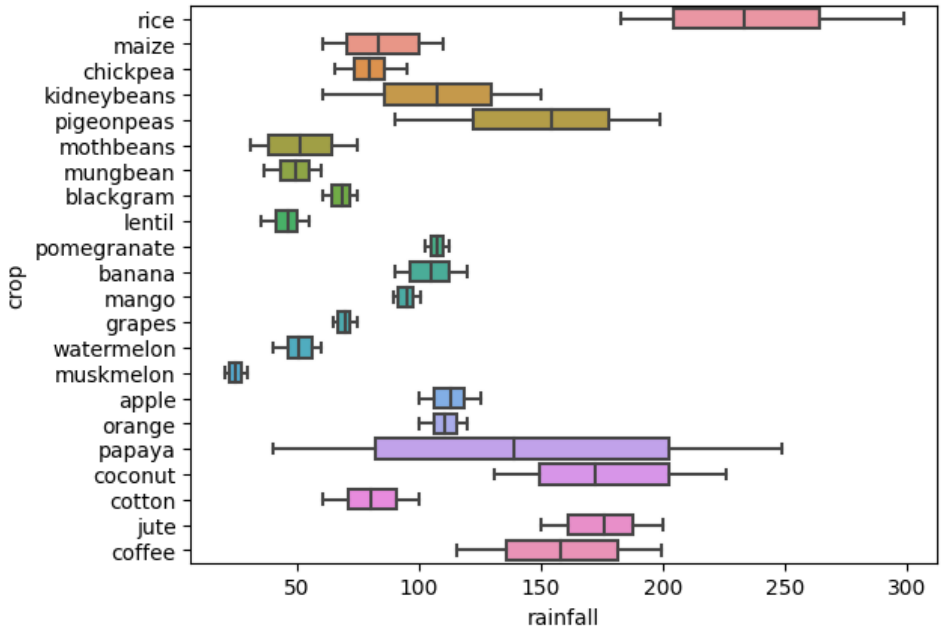
Figura 13: Relação entre rótulo e pH



Fonte: Autoria própria

O padrão se repete com o atributo *pH*. A maior parte dos valores está concentrada ao redor da média. Neste caso, tal resultado é previsível, visto que essa medida varia entre 0 e 14, e 7 representa um meio neutro (a média dos valores foi de aproximadamente 6,4)

Figura 14: Relação entre rótulo e precipitação



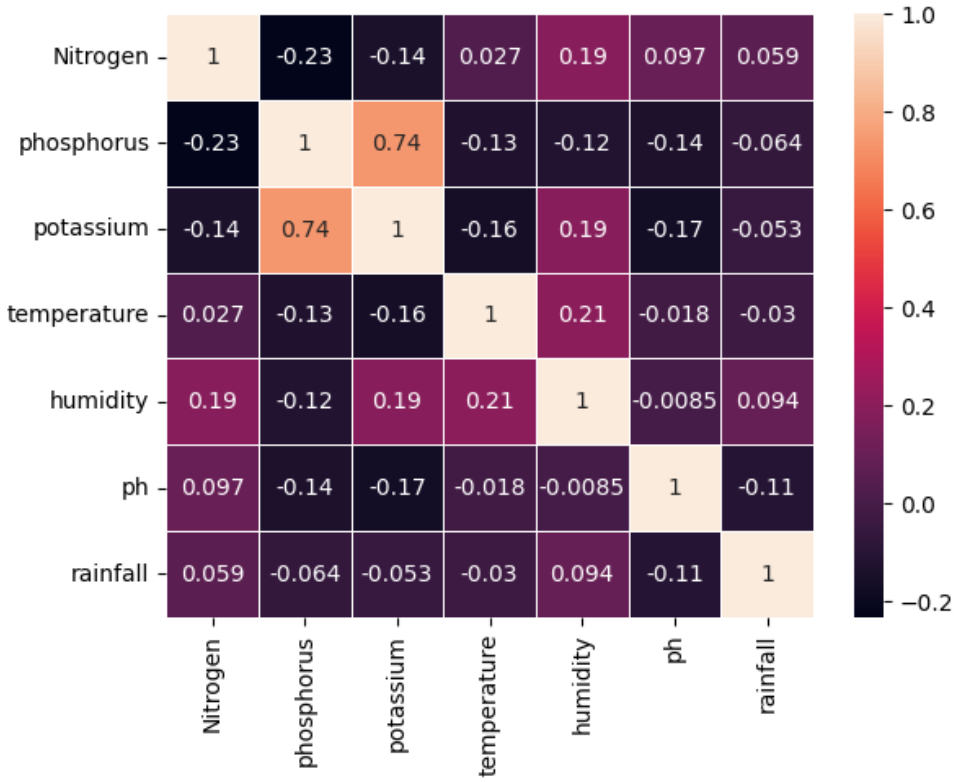
Fonte: Autoria própria

É possível perceber, então, que há certa separação no que tange aos atributos para cada tipo de cultura. Há indícios, portanto, de que é possível classificar o rótulo de novas observações a partir deste conjunto de variáveis.

3.3 Correlação entre as variáveis

Na figura 15 está representada através de um mapa de calor a matriz de correlações das variáveis:

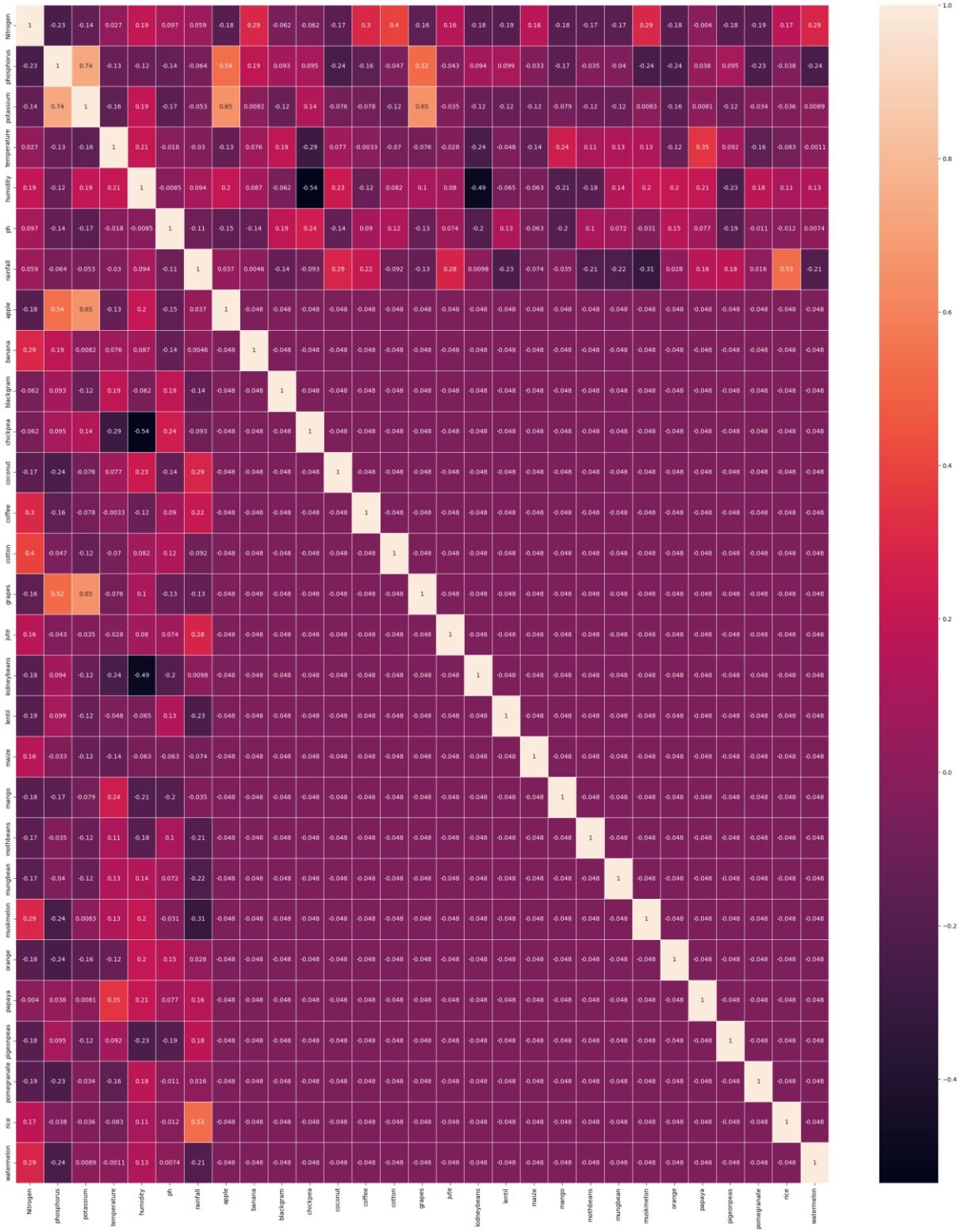
Figura 15: Mapa de calor da matriz de correlações



Fonte: Autoria própria

A maior parte das variáveis **não** está relacionada entre si, com exceção dos atributos *potassium* e *phosphorus*, que possuem uma correlação positiva considerável. Para que a classificação dos dados seja mais efetiva, um dos atributos pode ser removido. A figura 16 apresenta uma matriz de dispersão entre todos as culturas possíveis (os rótulos) e os atributos, o que é necessário para avaliar qual das duas variáveis relacionadas poderia ser removida:

Figura 16: Matriz de dispersão das variáveis



Fonte: Autoria própria

Sendo assim, uma possível saída para o problema seria remover a variável *phosporus* do conjunto de dados, visto que *potassium* tem maior relação com o atributo-alvo.

4 Referências Bibliográficas