**FA 084 – Introdução à Mineração de Dados**

**AP 522 –Mineração de Dados**

**ATIVIDADE PRÁTICA – AULA 10**

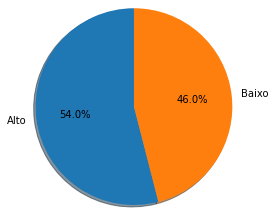
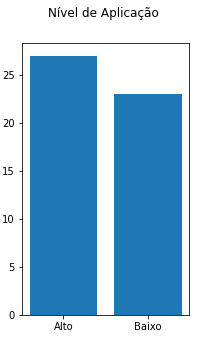
Gustavo Boaventura Cruz, 169227

1. **Questão A - Análise exploratória dos dados.**

Para a análise dos atributos dispostos no arquivo dado, foram sintetizados dois gráficos, sendo estes, de barras, onde permite uma melhor análise visual entre os valores e, um gráfico de setores (pizza), onde a proporcionalidade e frequência são facilmentes expostos.

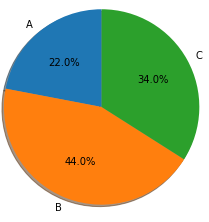
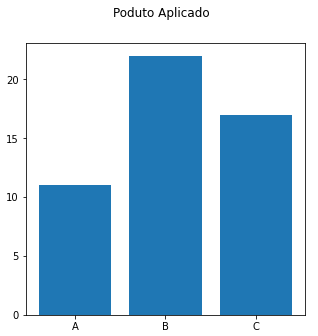
* **Nível de Aplicação de um produto (defensivo agrícola).**

Foram consideradas duas variáveis para o nível de aplicação de um produto, sendo estas, Alto e Baixo. Com a análise, foi observado o índice Baixo predominante em relação ao Alto (cerca de 8%), sendo respectivamente, 27 e 23.

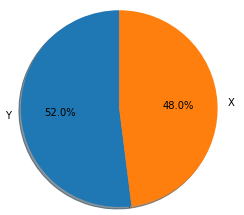


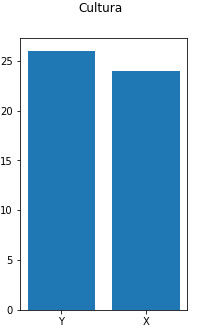
* **Produto (defensivo agrícola) aplicado.**

Foram consideradas três variáveis para o produto aplicado, sendo estas, A. B e C. Com a análise, foi observado a predominância do produto B em relação a C e A (cerca de 10% em relação a C e 22% para A), sendo respectivamente, 22, 17 e 11.



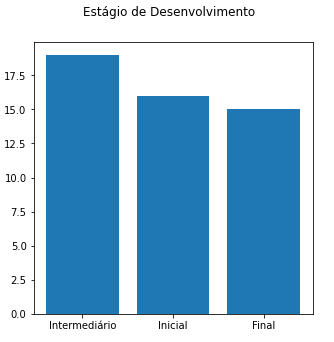
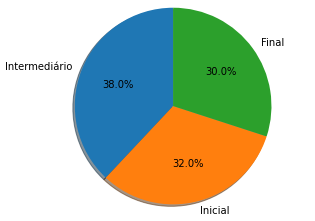
* **Cultura em que o produto foi aplicado.**

Foram consideradas duas variáveis para a cultura em que o produto foi aplicado, sendo estas, X e Y. Com a análise, foi observado a predominância da cultura Y em relação a X (cerca de 4%), sendo respectivamente, 26 e 24.

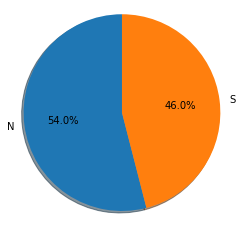
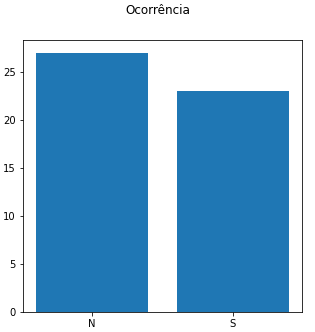


* **Estágio de desenvolvimento da cultura, quando o produto foi aplicado.**

Foram consideradas três variáveis para o desenvolvimento da cultura quando o produto foi aplicado, sendo estas, Inicial, Intermediário e Final. Com a análise, foi observado a predominância do desenvolvimento Intermediário em relação a Inicial e Final (cerca de 6% para a Inicial e 8% para a Final), sendo respectivamente, 19,16 e 15.



* **Indicador de ocorrência da doença.**

Foram considerados dois indicadores de ocorrência de doença, sendo estas, S e N. Com a análise, foi observado a predominância da cultura N em relação a S (cerca de 8%), sendo respectivamente, 27 e 23.

1. **Questão B - Separando os dados em conjunto de Treino e Teste.**

Para separar o conjunto Ocorrência, que foi o objetivo desta análise, foi usado uma proporção de 70/30 para treino e teste, respectivamente e, seed de 2020.

Ao início foi feita uma codificação dos dados, onde “Sim” são transformados em “1” e “Não” em “0”. Além de facilitar a manipulação destes, também foi necessário para utilizar o KNN (Item 3), donde para usa-lo, teníamos que transformar object em float. Para isto foi usado OneHotEncoder da biblioteca do Sklearn.

Após a separação do conjunto Ocorrência em Treino e Teste, tivemos os seguintes tamanhos de variáveis X\_train e X\_teste de tamanho (35, 8) (15, 8) respectivamente (onde 35 representa 70% dos conjuntos dos dados e 15 30%). Para y\_train e y\_test, foram dados os mesmos 35 e 15 dados aleatórios, sendo eles 17 e 6 com ocorrência de doença, respectivamente.

1. **Questão C - Técnica KNN (K Nearest Neighbor).**

A técnica KNN novamente do Sklearn, foi utilizando cinco valores distintos para K, sendo 1, 3, 5, 7 e 9. Sendo construído pelo código demonstrado abaixo.

knn1.fit(X\_train, y\_train)

knn3.fit(X\_train, y\_train)

knn5.fit(X\_train, y\_train)

knn7.fit(X\_train, y\_train)

knn9.fit(X\_train, y\_train)

1. **Questão D - Predicação.**

Após a construção dos respectivos KNN, foi feita a predição destes, separando os testes dos treinos (X\_test e X\_train), novamente demonstrado pelo código abaixo.

y\_pred1 = knn1.predict(X\_test) y\_pred11 = knn1.predict(X\_train)

y\_pred3 = knn3.predict(X\_test) y\_pred33 = knn3.predict(X\_train)

y\_pred5 = knn5.predict(X\_test) y\_pred55 = knn5.predict(X\_train)

y\_pred7 = knn7.predict(X\_test) y\_pred77 = knn7.predict(X\_train)

y\_pred9 = knn9.predict(X\_test) y\_pred99 = knn9.predict(X\_train)

1. **Questão E - Acurácias obtidas em Treino e Testes.**

Foram feitas o total de 10 acurácias para os modelos, cinco para os testes e cinco para os treinos. Obtidas a partir do código metrics.accuracy\_score, os resultados são demonstrados abaixo.

Accuracy Train k=1: 1.0

Accuracy Train k=3: 0.9142857142857143

Accuracy Train k=5: 0.8285714285714286

Accuracy Train k=7: 0.8

Accuracy Train k=9: 0.8

Accuracy Test k=1: 0.8

Accuracy Test k=3: 0.5333333333333333

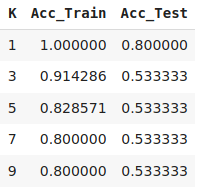
Accuracy Test k=5: 0.5333333333333333

Accuracy Test k=7: 0.5333333333333333

Accuracy Test k=9: 0.5333333333333333

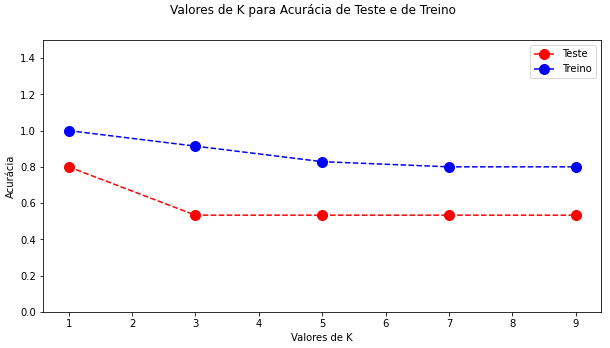
1. **Questão F - Tabela com as Acurácias.**

Com os resultados das acurácias de teste e treino obtidos no item anterior, foi sintetizado uma tabela, onde pode se analisar de uma forma mais clara os valores de cada acurácia para os respectivos valores de K.

****

1. **Questão G - Construção do gráfico (K x Acurácia).**

Foi sintetizado um gráfico com os valores de cada acurácia, sendo elas de teste e treino, como demonstrado os valores no item 6, para os seus respectivos K.



A acurácia de treino é demonstrada pela linha azul e a de teste pela vermelha. Foi escolhido este estilo de gráfico, já que se torna mais visível a diferença entre as duas acurácias, e em cada ponto K está demonstrado por um ponto cheio, onde mostra exatamente o valor de cada acurácia.

Podemos notar a diferença e similaridade entre estas. A acurácia treino se manteve para todos os valores de K superior a de teste, onde para o valor de K = 7, elas se mantiveram constantes, com 0.8 e 0.533 respectivamente. Podemos notar também a acurácia para K=1 no conjunto de treinamento foi 1, ou seja, 100%.