Data Scientist (Graduate Trainee) Pre-Screening Assignment

Name: Yehezkiel Edininta Simanjuntak

**Question No. 1**

A customer informed their consultant that they have developed several formulations of petrol that gives different characteristics of burning pattern. The formulations are obtaining by adding varying levels of additives that, for example, prevent engine knocking, gum prevention, stability in storage, and etc. However, a third party certification organisation would like to verify if the formulations are significantly different, and request for both physical and statistical proof. Since the formulations are confidential information, they are not named in the dataset.

1. A descriptive analysis of the additives (columns named as “a” to “i”), which must include summaries of findings (parametric/non-parametric). Correlation and ANOVA, if applicable, is a must.
2. A graphical analysis of the additives, including a distribution study.
3. A clustering test of your choice (unsupervised learning), to determine the distinctive number of formulations present in the dataset.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

First, we import the data. From the picture above, we can see that there are 9 columns and each column has 214 rows with a float data type and no NULL values.

Graphical user interface, table

Description automatically generated

Descriptive statistical analysis helps to describe basic features of a dataset and obtains a short summary about the sample and measures of the data. One way in which we can do this is by using the describe() function in pandas. Using the describe function and applying it on your dataframe, the "describe" function automatically computes basic statistics for all numerical variables. It shows the mean, the total number of data points, the standard deviation, the quartiles and the extreme values. This function will give you a clearer idea of the distribution of your different variables.

Graphical user interface, application

Description automatically generated

One way to measure the strength of the correlation between continuous numerical variables is by using a method called Pearson Correlation. For the correlation coefficient, a value close to 1 implies a large positive correlation, while a value close to -1 implies a large negative correlation and a value close to 0 implies no correlation between the variables. Pearson Correlation is the default method of the function "corr". Like before, we can calculate the Pearson Correlation of the of the 'int64' or 'float64' variables.

From the image above, for example we can see that variable “a” and “g” have a positive correlation, while variable “a” and “e” have a negative correlation.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

A box-plot is a very useful and standardized way of displaying the distribution of data based on a five-number summary (minimum, first quartile, second quartile (median), third quartile, maximum). It helps in understanding these parameters of the distribution of data and is extremely helpful in detecting outliers. The box plot shows that there are outliers in each column. However, because the outlier is not too extreme and has a value close to the whisker, we can leave it as valid data.

Graphical user interface

Description automatically generated

The number of clusters that we choose for the algorithm shouldn’t be random. Each and Every cluster is formed by calculating and comparing the mean distances of each data points within a cluster from its centroid.

We can choose the right number of clusters with the help of the Within-Cluster-Sum-of-Squares (WCSS) method. WCSS Stands for the sum of the squares of distances of the data points in each and every cluster from its centroid. The main idea is to minimize the distance between the data points and the centroid of the clusters. The process is iterated until we reach a minimum value for the sum of distances. We can use the elbow method to visualize the inertia for different K values.

To find the optimal value of clusters, the elbow method follows the below steps:

1. Execute the K-means clustering on a given dataset for different K values (ranging from 1-10).
2. For each value of K, calculates the WCSS value.
3. Plots a graph/curve between WCSS values and the respective number of clusters K.
4. The sharp point of bend or a point (looking like an elbow joint) of the plot like an arm, will be considered as the best/optimal value of K.

Based on the picture above, we can see that the most optimum number of clusters is 3 (good value for K) because at that point it looks like an elbow joint.

Calendar

Description automatically generated

Table

Description automatically generated with medium confidence

Chart

Description automatically generated with medium confidence

The last step is to create a model using the K-Means method and visualize the clustering process using a scatter plot. From the picture above, we can see that the three clusters are quite evenly separated, and there is also a distance between the centroids. With these results we can conclude that the number of formulations present in the dataset is 3.

**Question No. 2**

A team of plantation planners are concerned about the yield of oil palm trees, which seems to fluctuate. They have collected a set of data and needed help in analysing on how external factors influence fresh fruit bunch (FFB) yield. Some experts are of opinion that the flowering of oil palm tree determines the FFB yield, and are linked to the external factors. Perform the analysis, which requires some study on the background of oil palm tree physiology.

Table

Description automatically generated

A picture containing graphical user interface

Description automatically generated

This time we’ll visualize any existing variable with target output FFB\_Yield using a scatter plot and an added linear line called a “regression line”, which indicates the relationship between the two (SoilMoisture, Average\_Temp, Min\_Temp, Max\_Temp, Precipitation, Working\_days, HA\_Harvested with the FFB\_Yield variable). The main goal of this plot is to see whether the existing variable has any impact on the FFB\_Yield.

Graphical user interface

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Graphical user interface, chart

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Chart

Description automatically generated

Graphical user interface, application

Description automatically generated

From the visualization above, we can see that the data points of the **SoilMoisture** and **Average\_Temp** variables have a great variation and the regression line is almost horizontal. We can’t determine what form of correlation we can expect but assuming there is a linear relationship. A nearly flat regression line indicates that the correlation coefficient is around 0. **Min\_Temp** and **Working\_days** have a slight positive correlation with FFB\_Yield, while **Max\_Temp** has a slight negative correlation. For the **Precipitation** variable, the visualization we get is an increasing gradient of the regression line which shows us that we can expect a **POSITIVE** correlation. This means that when one variable increases, the other variable also increases. As for the **HA\_Harvested** variable, the visualization we get is a decreasing regression line indicating a **NEGATIVE** correlation.

A picture containing table

Description automatically generated

To answer this second question, the first technique used is Pearson Correlation. The Pearson correlation method, will give you two values, the Correlation coefficient and the p-value. For the correlation coefficient, a value close to 1 implies a large positive correlation, while a value close to -1 implies a large negative correlation and a value close to 0 implies no correlation between the variables. Next, the p-value will tell us how certain we are about the correlation that we calculated. For the p-value, a value less than 0.001 gives us a strong certainty about the correlation coefficient that we calculated. A value between 0.001 and 0.05 gives us moderate certainty, a value between 0.05 and 0.1 will give us a weak certainty and a p-value larger than 0.1 will give us no certainty of correlation at all.

From the picture above, we can see that there is only one variable, namely **Precipitation** which has a P-value smaller than 0.001 so it can be said that there is **strong evidence** that the correlation is **SIGNIFICANT**.

The second technique used here is Feature Importance. Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction. Feature importance scores can be calculated for problems that involve predicting a numerical value, called regression, and those problems that involve predicting a class label, called classification.

The scores are useful and can be used in a range of situations in a predictive modeling problem, such as:

* Better understanding the data.
* Better understanding a model.
* Reducing the number of input features.

Feature importance scores can provide insight into the dataset. The relative scores can highlight which features may be most relevant to the target, and the converse, which features are the least relevant. This may be interpreted by a domain expert and could be used as the basis for gathering more or different data.

Feature importance scores can provide insight into the model. Most importance scores are calculated by a predictive model that has been fit on the dataset. Inspecting the importance score provides insight into that specific model and which features are the most important and least important to the model when making a prediction. This is a type of model interpretation that can be performed for those models that support it.

Feature importance can be used to improve a predictive model. This can be achieved by using the importance scores to select those features to delete (lowest scores) or those features to keep (highest scores). This is a type of feature selection and can simplify the problem that is being modeled, speed up the modeling process (deleting features is called dimensionality reduction), and in some cases, improve the performance of the model.

**Graphical user interface, text, application

Description automatically generated**

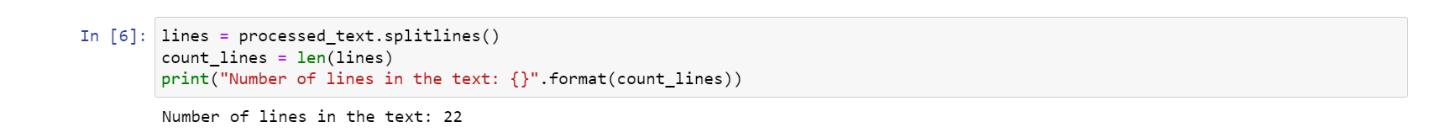
From this we can see that the variables that most influence the FFB\_Yield variable are the **Precipitation** and **HA\_Harvested** variables.

**Question No. 3**

Feed the following paragraph into your favourite data analytics tool, and answer the following;

1. What is the probability of the word “data” occurring in each line ?
2. What is the distribution of distinct word counts across all the lines ?
3. What is the probability of the word “analytics” occurring after the word “data” ?

The first step is to enter text and store it in a variable. After that, preprocessing is carried out such as case folding to convert the text into lowercase letters, correcting some words such as it's to it is and finally removing punctuation marks.



The text that has been entered and has been preprocessed consists of 22 lines.

Text, letter

Description automatically generated

After that, the word tokenization process is carried out in which the entire text will be split into a single word. The number of words contained in the text is 318 words.

Graphical user interface, text, application

Description automatically generated

After we get the words contained in the text, then we can count the number of word “data” contained in the text which is 18. To calculate the probability of the word “data” appearing in each line, we can divide the number of word “data” the text by the number of lines in the text, in this case 18/22 which is equals to 0.82. So the probability of the word “data” appearing in each line is **0.8181** or **81.81%**.

Graphical user interface, text

Description automatically generated with medium confidence

Chart, bar chart, histogram

Description automatically generated

To be able to see how many words are unique in each line, we need to find the words that are in each line and then display the unique ones with the set function. From the picture above, we can see that the number of unique words in each line is almost the same, only one line, namely the last line which has the least number of unique words, which is 5.

Graphical user interface, text, application, email

Description automatically generated

For question number 3 part C we can determine the probability of the word “analytics” occurring after the word “data” with the help of n-grams. An N-gram means a sequence of N words. So for example, “Medium blog” is a 2-gram (a bigram), “A Medium blog post” is a 4-gram, and “Write on Medium” is a 3-gram (trigram). For this problem we need a bigram model. Now because this is a bigram model, the model will learn the occurrence of every two words, to determine the probability of a word occurring after a certain word.

We’re calculating the probability of word “analytics” occurring after the word “data” then the formula for this is as follows:

which is the number of times the words occurs in the required sequence, divided by the number of the times the word before the expected word occurs in the whole text. So in conclusion we get a probability of **0.3333** or **33.33%**.