Coursework Report

MACHINE LEARNING – ST3189

STUDENT ID - 200699684

CONTENTS

1. Unsupervised Learning

Refers to recognizing patterns within data that has not been labeled before

Clustering – Clustering is the process of identifying groups in the dataset where the data points in one group have similar characteristics while different to the data points in other groups.

Literature Review

According to (Andreinna , 2020) article, it explains that 'the optimal number of clusters that fit to this dataset is 3 .Also as per (Mishra, 2018) ,lower level of alcohol has a high chance of being the wine from class 2. Based on these facts, research questions were formed.

Research Questions.

1.Can there be 2 clusters in the given dataset?

2.Can there be 4 clusters in the given dataset?

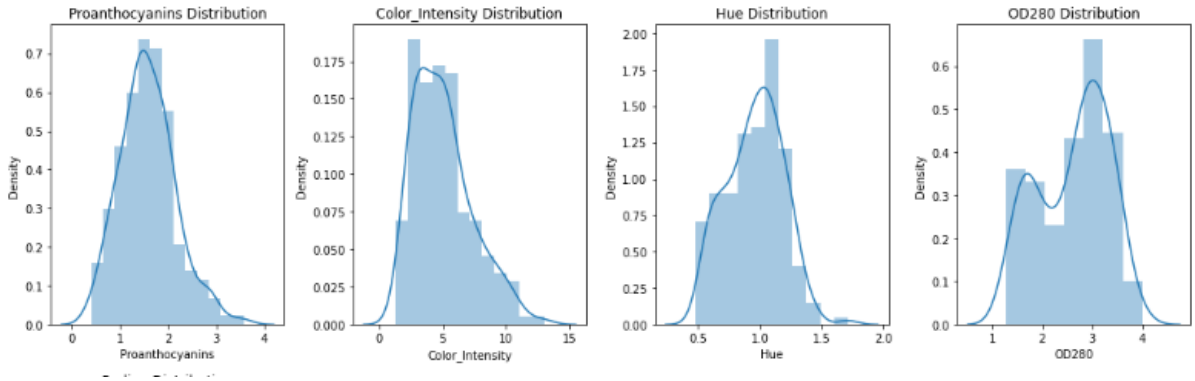
3. Does actually lower level of alcohol help to categorize the dataset into clusters?

Introduction to the Analysis –

The Dataset is about results of a chemical analysis of wine' (cite). Thirteen variables were given in order to conduct the analysis.

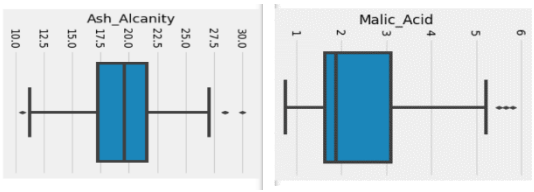
1. Exploratory Data Analysis.

The null values of the data set were dropped as the first step. Then histogram plots were drawn in order to understand whether the variables have been normally distributed or not.

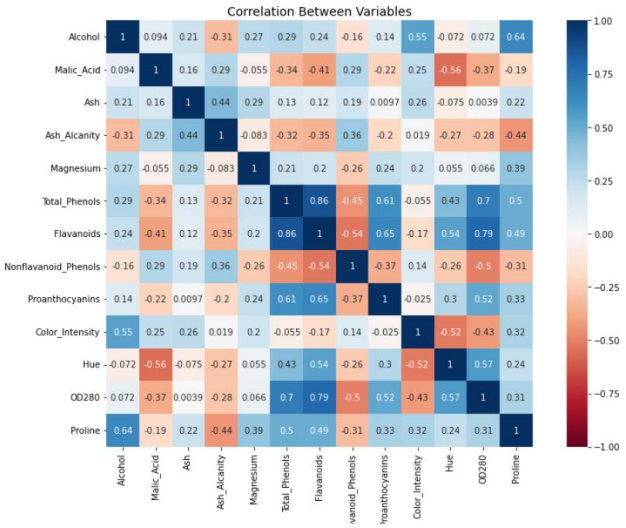


The graph shows that 'Proanthocyannis' variable seems to follow a normal distribution but 'OD280' does not seem to follow a normal distribution as it has two spikes. The plots for other variables are done in the submitted Jupyter Note book.

Then boxplots plots were drawn to identify whether outliers do exist or not. Outliers were present in some variables as shown in the following boxplots.



Therefore the outliers were removed from the dataset in order to make sure that the findings from the model are accurate and valid. Then a correlation coefficient matrix was plotted in order to find whether there exists a linear relationship among the variables, if there exists, to remove those variables in order to make sure the model is accurate.(To prevent from multicollinearity)



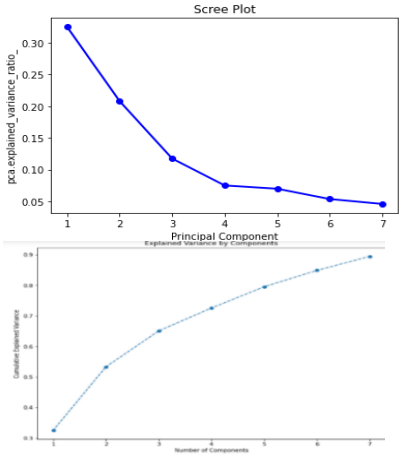
The plot shows that ‘Flavanoids’ variable seems to have significant correlation values with ‘Total_Phenols’ (0.86), and with 0D280 (0.79), therefore it was identified as a variable with high correlation coefficient, therefore it was removed from the dataset.

After standardizing the data to ensure that data are not biassed, Principal Component analysis was conducted in order to reduce the number of dimensions of the data set.

Principal Component Analysis (PCA)

“PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets” (Pierre, 2022). A scree plot was plotted inorder to identify the number of Principal Components required for the model. The x axis is the number of Principal Components and the y axis is the ratio of the explained variance of each principal component. The cumulative graph shows the cumulative value as the y axis.

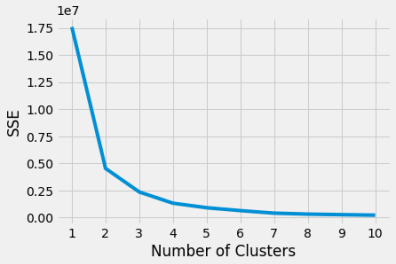
Elbow Method – We use this method in order to clarify the number of principal components required for the model by looking at the scree plot. We check for the elbow point where the curve becomes flat afterwards . By looking at the Scree Plot below, we can see that number 4 in x axis is the elbow point, therefore 4 Principal Components were sufficient for the model.

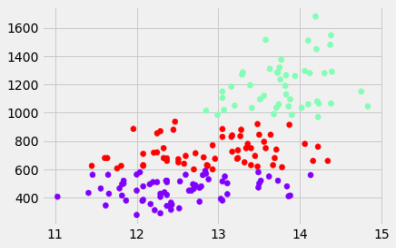


K- Means Clustering

"The algorithm works iteratively to assign each data point to one of k groups based on the features that are provided" (Trevino, 2016).

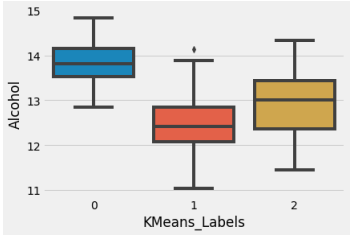
Inertia –. "It is the measure of intra-cluster distances, which means how far away the datapoint is concerning its centroid" (Bhavsar, Srivastava, & Awasthi, 2020).





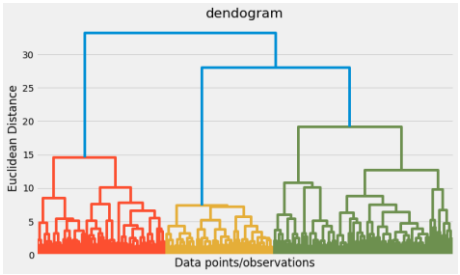
As per the Elbow method, we can say that 3 clusters would be sufficient for the model.

The below scatter plot shows how the data points were clustered into two where 'Alcohol','Proline' being x axis and y axis. It shows that lower level of Proline tends to fall into one cluster but lower level of Alcohol separate the data set into two clusters not one. Also boxplots were drawn in order to understand how the model has divided the dataset into 3 clusters for each variable, as an example, the boxplot plotted for Alcohol clarifies what was identified in scatterplot while alcohol being the x axis.

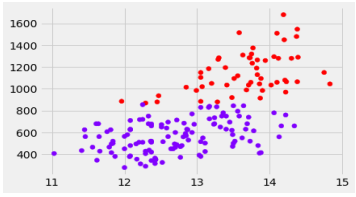


Hierachical Clustering–

'Starting from treating each data point as a single cluster, model iteratively processes to merge data points and make a cluster' (Bock, 2022). Clusters are merged based on the distance between them and to calculate the distance between the clusters we have different types of linkage methods. The plot we use to visualize the hierarchy

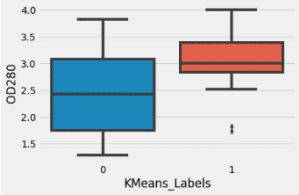


The figure on left shows how the model has identified three clusters starting from each data point being categorized as one cluster (From bottom to top)



Answer to Question 1. – K means clustering model was fitted as K is equal to 2 and the findings were plotted below. By looking at these plots we can say that 2 clusters also fit the model, but as per the scree plot, 3 clusters is the optimal solution for this model

Question 2 – Simply by looking at the scree plot, we can say that 3 clusters almost explain the variation when 5 clusters were used, therefore 5 is not a good number for clusters for this model



Question 3- When looking at the boxplot graph included above, we can identify that lower level of alcohol does fall into two clusters, not into one. But very lower level of Alcohol ultimately falls into one cluster.

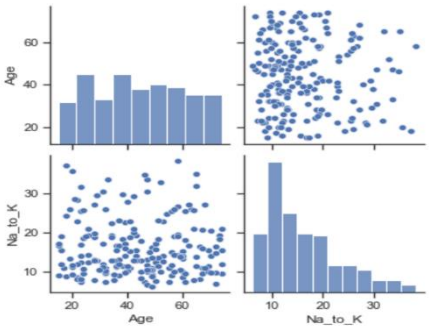
2. Supervised Learning - Classification

“The model tries to predict the correct label of a given input data” (Keita, 2022).

The dataset is about predicting the accurate drug type for a patient given the person’s Age,Sex, BP, Cholesterol Level, Na to K ration.

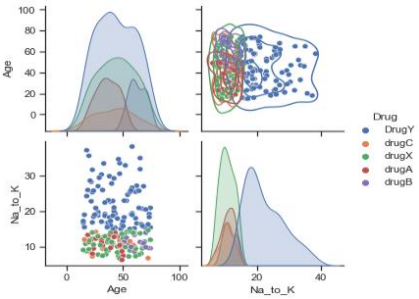
Firstly, the null values were checked, as there were no null values, Exploratory Data Analysis was conducted.

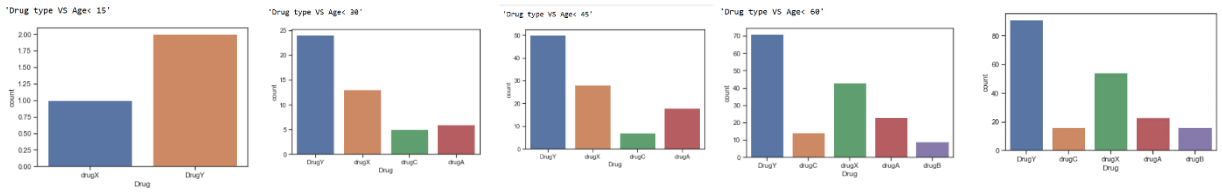
Exploratory Data Analysis



The upper left figure is the distribution of age, it does not seem to follow a normal distribution, while ‘Na to K’ is skewed to the right. Upper right and lower left graphs show the scatterplots of those two variables.

Figure on the right shows how each particular drug type has been positioned in each of the plots described above. Then boxplots were plotted for numerical variables in order to identify whether outliers do exist or not. The boxplots show that no outliers do exist (Included in the submitted note book)



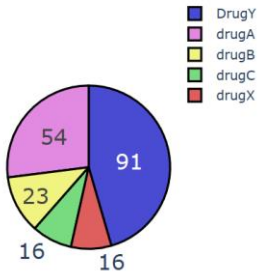


Age variable was categorized into 5 bins in order to identify how the drug type changes along with age category. It shows that for people who are aged less than 15 years have used only drugX and DrugY. Drug B has been used by people who are aged 60 years and above only. Then the outliers were checked using boxplots

The correlation matrix was drawn as the next step in order to identify whether there are correlated variables in the dataset. As -0.063 is close to zero,we can conclude that there does not exist alinear relationship among the variables in the dataset.

Fitting the model - In order to fit the model, the categorical variables were transformed as dummy variables using 'label encoder'method, then the dataset was split as dependent and independent variables while 'Drug' type being the dependent variable. Afterwards the variables were standardized using standard scaler.

The following pie chart was plotted in order to understand the count for each drug type has been.



As we have to create a model and to validate results, the dataset was split into test and train sets with a ratio of 0.3(which means 0.7 proportion was for the train set and 0.3 for the test set)

K Nearest Neighbors Classification

"Classified by "MAJORITY VOTES" for its neighbor classes" (Kang). The number for K has to be given manually to the model, which are the number of categories in the dependent variable

Random Forrest Classification

"It builds decision trees on different samples and takes their majority vote for classification and average in case of regression"(Sruthi ER, 2023).

XG Booster Classifier

This module checks whether the created decision tress are being overfited or not

After fitting these models for the dataset, Train Score, Test Score, Accuracy Score were calculated in order to compare and find the best model that fits the dataset.

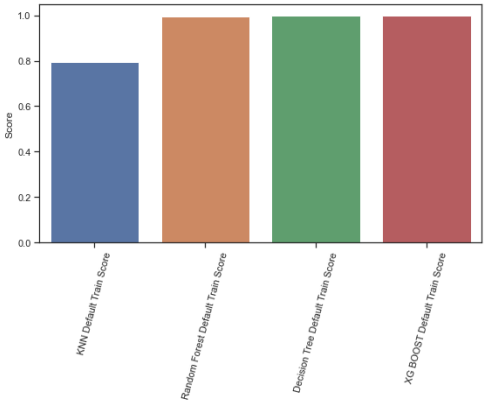
Train Score – This explains how the model was fitted in the training data, if the model has fitted so well in the data with high variance then it causes overfitting, hence a low train score value

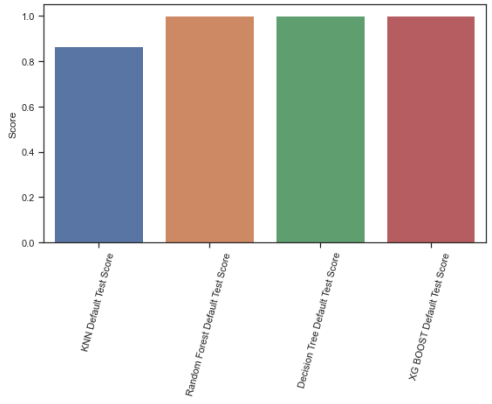
Test Score – This explains how well the model is generalized in the test set, higher the score value, better the model is.

	Score
KNN Default Test Score	0.866667
Random Forest Default Test Score	1.000000
Decision Tree Default Test Score	1.000000
XG BOOST Default Test Score	1.000000

	Score
KNN Default Train Score	0.792857
Random Forest Default Train Score	0.992857
Decision Tree Default Train Score	0.997521
XG BOOST Default Train Score	0.997521

Except KNN model, all the other models have taken the Test Score as 1, practically this would not happen but as due to the dataset selected this kind of a result would not bring doubts, as the train scores for the models are lesser than 1 and they are different to one another.





According to (Kohli, 2019),

Precision - Describes the level of your predictions for each variable as a fraction.

Recall – The level of caught positive cases as a fraction.

F1 value –The number of correct positive predictions as a fraction

Classification reports for each of the model has been created and included in the submitted document, as per those reports we can identify that Random Forrest classifier is the best model.

Regression

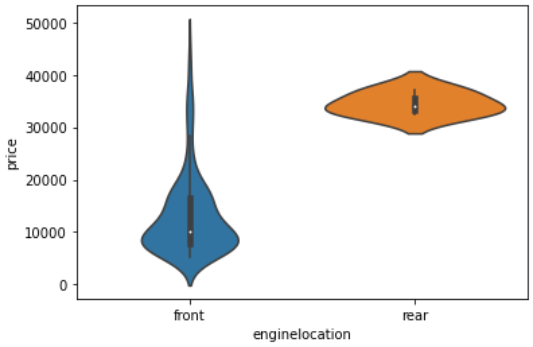
Literature Review

(Iakubovskyi, 2023) article states that diesel cars are more expensive than petrol vehicles and the percentage is 15%.

Research Question

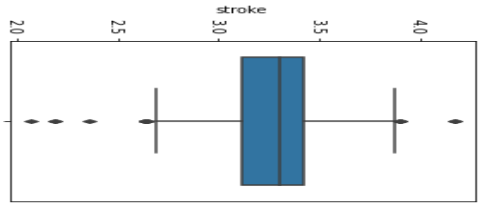
Does actually diesel cars value more than petrol cars?

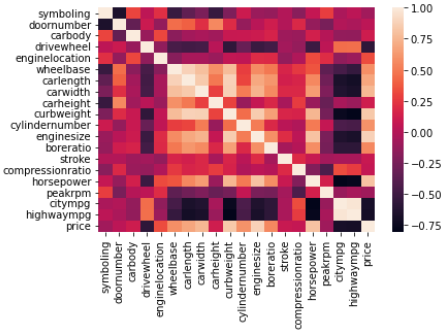
As there were no null values in the dataset when checked, exploratory data analysis was conducted. Firstly, the data features were categorized as dependent variables and independent variables. Then features which were not important such as ‘Car_ ID’ were removed from the dataset. Then the duplicate car names were converted into the standard names. Thereafter, violin plots were plotted in order to understand the distributions of different feature variables.



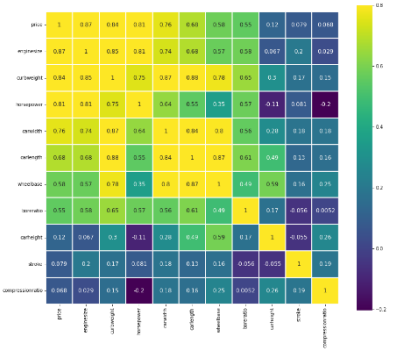
As an example, using the attached violin plot, we can realize that the cars which the engine has been located in the rear seems to have a higher Car Price in general compared to the cars which the engine has been located in the front, and they seem to follow a distribution which is skewed to the right. Likewise violin plots were drawn for other variables as well. As boxplot is also included in the violin plot, existence of outliers was also checked through that. But in order to get a clearer picture of the outliers, separate boxplots were drawn as well

.As an example, by looking at this boxplot which was drawn for variable stroke, we can identify that this variable possesses outliers. Outliers were removed from the dataset afterwards.





As the next step, the Correlation Coefficient Matrix and the heat map was drawn inorder to check the correlation values among the feature variables of the dataset. Existence of Correlation means Multicollinearity does exist in the data set which is a violation of one of the main assumptions of linear regression which means that the error terms of the model does not follow a normal distribution, which ultimately affects the validity of the model.



So as per these graphs and codes 'carlength', 'curbweight', 'enginesi e', 'highwaympg' variable detected as correlated variables as they se em
to have correlation coefficient values greater than 0.85

Then all the categorical variables were converted into numerical one s
using Dummy variables.

Fitting the Model

Multiple Linear Regression- In order to fit the model and come up with results, the data set had to split into two as train set and test set, then model coefficients, predicted values for the dependent variable and intercept values were calculated. Assumptions of the model.

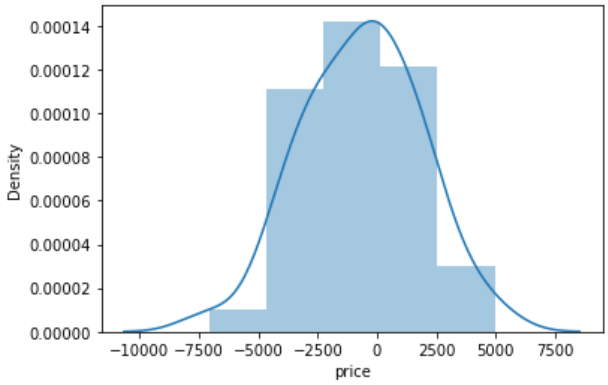
So the equation of the model is:

Yi = B0* +B1*symbolling + B2*fueltype + B3*aspiration + B4*doornumber+ B5*carbody + B6*drivewheel + B7*enginelocation + B8*wheelbase + B9*carlength +B10*carwidth + B11*carheight + B12*curbweight + B13*cylindernumber +B14*enginesize + B15*fuelsystem +B17*boreratio +B18*stroke + B19*compressionratio + B20*horsepower +B21*peakrpm +B22*citympg + B23*highwaympgl + Errori

Where B0 is the intercept value and B1, B2,,, B23 are the coefficient values for each variable and Errori s are residuals

The following assumptions were made prior to fit the model into the dataset.
1.There exists a linear relationship between the dependent variable and independent variables.
2.Error terms have constant variance
3.Error terms are normally distributed
4.There exist no linear relationship among independent variables

Then the residuals of the model were plotted using a histogram in order to clarify that residuals were normally distributed, if not the linear regression model is invalid. As per the graph, we can conclude that the residuals were normally distributed.



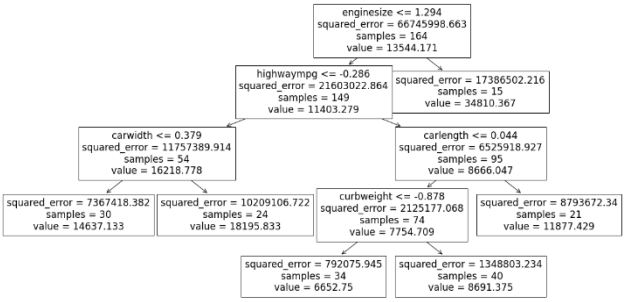
The B0,B1,,B23 values were calculated(Included in the JNotebook)

R squared value was 0.885, which means 88.5% of the total variation of the dependent variable is explained by the model, which is a good value. The R.M.S.E (Root Mean Squared Error) value for this model is 44.9

RMSE measures the average difference between the predicted values and the original values.

Decision Tree Regressor

According to the diagram, we can identify that the first decision was taken on engine size value of 1.294,less than that amount falls into one leaf and greater than into the other leaf. Second decision node was based on. Highwaympg variable.



Random Forrest Regressor model was also fitted for the dataset. The results of R squared value and RMSE value were calculated

Model Name	Multiple Linear Regression	Decision Tree Regressor	Random Forrest Regressor
R Squared	0.885	0.885	0.885
RMSE	44.9	50.759	37.789

Answer to the research Question – Using Multiple Regression model, we were able to find the coefficient of fuel type as -1937.333672.

Value one was indicated as gas (petrol) and zero was indicated as diesel. Therefore assuming that fueltype variable is significant, compared to a diesel car, a petrol one reduces the price by 1937.3336 amount of units of that particular currency.

Conclusion – Comparing the three models fitted to the dataset, as R squared values are same for all the models, Random Forrest Regressor will be chosen as the best model as its RMSE value is lesser than the other two models.

Bibliography

Andreinna, S. (2020, July 14). K-Nearest Neighbors (KNN) with wine dataset. Retrieved April 1, 2023, from https://medium.com/@shaulaandreinnaa/k-nearest-neighbors-knn-with-wine-dataset-3794acaec833

Bhavsar, S., Srivastava, U., & Awasthi, S. (2020, December 03). Understanding K-means clustering. Retrieved April 1, 2023, from https://datamahadev.com/understanding-k-means-clustering/

Bock, T. (2022, September 13). What is hierarchical clustering? Retrieved April 1, 2023, from https://www.displayr.com/what-is-hierarchical-clustering/

Https://www.saedsayad.com/decision_tree_reg.htm. (n.d.). Retrieved April 2, 2023, from https://www.saedsayad.com/decision_tree_reg.htm

Iakubovskyi, D. (2023, March 08). Which factors form the used car price? Retrieved April 2, 2023, from https://medium.com/@dima806/which-factors-form-the-used-car-price-shap-values-based-on-the-kaggle-vehicle-dataset-33b0a2f7896c

Kang, M. (n.d.). K-nearest neighbor algorithm - University of Nevada, Las Vegas. Retrieved April 1, 2023, from https://www.mkang.faculty.unlv.edu/teaching/CS489_689/05.KNN.pdf

Keita, Z. (2022, September 21). Classification in machine learning: A guide for beginners. Retrieved April 1, 2023, from https://www.datacamp.com/blog/classification-machine-learning

Kohli, S. (2019, November 18). Understanding a classification report for your machine learning model. Retrieved April 2, 2023, from https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Mishra, N. (2018, December 11). FN(UCI wine dataset). Retrieved April 1, 2023, from https://medium.com/intuitions/fn-uci-wine-dataset-ef9f5dfc20c4

Pierre, S. (2022, August 08). A step-by-step explanation of principal component analysis (PCA). Retrieved April 1, 2023, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Sruthi ER. (2023, March 24). Understand random forest algorithms with examples (updated 2023). Retrieved April 2, 2023, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Trevino, A. (2016, December 6). Introduction to K-means Clustering. Retrieved March 1, 2023, from https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering

What is the K-nearest neighbors algorithm? (n.d.). Retrieved April 1, 2023, from https://www.ibm.com/topics/knn

Links for the datasets

Regression - Car Price Prediction Multiple Linear Regression | Kaggle

Classification -Drug Classification | Kaggle

Clustering - Wine Dataset for Clustering | Kaggle