*CA2 Data Preparation and Machine Learning*

***Predicting the respondent diabetic using***

***different machine learning algorithm.***

*By*

*Rachel Mae Lademora – 2023396*

*Higher Diploma in Science in Data Analytics for Business*

*Strategic Thinking*

*James Garza*

*CCT College*

*Dublin, Ireland*

CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA2 Data Preparation and Machine Learning |
| **Lecturer Name:** | David McQuaid  Dr. Muhammad Iqbal |
| **Student Full Name:** | Rachel Mae Lademora |
| **Student Number:** | 2023396 |
|  |  |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** | January 02,2023 |
| **Date of Submission:** | January 02,2023 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Abstract**

One of the major popular topics in the United Stated is the health and nutrition of the people. We all know that many cases of imbalance nutrition United States have. Since there are a lot of cases of unhealthy lifestyle that they eat, there are many cases of different age that has the nutrition problem.

One example of the nutrition problem is obesity that can lead to a serious disease. One of the popular diseases in the United States that obesity has increase of the severe risk of having a health condition of high blood pressure, diabetic, stroke and heart disease. There are many cases of type 2 diabetic in the Unites States because there are many people there that eat unhealthy food that cause them to be overweight.

The report aim is to have a thorough determine the comparison which of the LDA and PCA is more accurately to be used to train the model. This will forecast by using KNN, Decision Tree Classifier, Random Forest and kmeans clustering.

Finally, the report will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to explain in detailed stages of Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation and Deployment.

Table of Contents

Business Understanding**1**

Data Understanding2

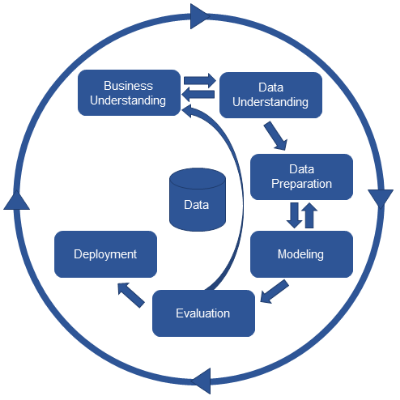
Data Preparation3

Modelling**4**

Evaluation5

Deployment6

References………………………………………………………………………



**Business Understanding**

Business understanding is the first phase of the CRISP-DM methodology. The main purpose is to well understand the business and its necessity, which is the objective and the requirements to use this project.

The principal objective of this project to predict the respondent that has diabetic in relation with their age and who undergo the blood glucose after fasting. This provide the comparison prediction which of the LDA and PCA is most accurately to be used in preparation of machine learning model. The hypothesis of this project is to compare the accuracy results of the machine learning models to determine which one provide the most efficient accuracy.

**The requirements of this project are:**

* Identify the features and target variable.
* Use the Exploratory Data Analysis (EDA) to understand the data.
* Apply machine learning models to choose one with a moderate accuracy score. To avoid the over fitting and under fitting of the models.
* Implement the chosen machine learning model to predict the respondent diabetic using the PCA to have the better accuracy score.

General Goal:

The project aim to asses some machine learning models to predict the respondent diabetic in classes features and compare supervised and unsupervised machine learning which of them will give the best efficient accuracy score.

The list of important tools and technologies that will be used in the project are:

* List of the Python Libraries:

1. Pandas
2. Numpay
3. Matplotlib
4. Seaborn

* Modelling: KNN, Decision Tree Classifier- a machine learning model to test using of branching of nodes, Random Forest, and kmeans clustering.

**Data Understanding**

Data understanding is the second phase of the CRISP-DM Methodology, it means to know about what I have understand to the data.

In this second stage of the CRISP-DM it is crucial to take some time to look every detail of the project data that is stated on the dataset. In order to avoid and encounter some errors when I will proceed to the data preparation. Data preparation is the vital part of the data analysis which is on the third phase of the CRISP-DM methodology. On that part I need to execute the data analysis like cleaning the data which vital before performing the modelling.

The dataset is based on the National Health and Nutrition Examination Survey (NHANES). administered by the Canters for Disease Control and Prevention (CDC), collects extensive health and nutritional information from a diverse U.S. population. (archive.ics.uci.edu, n.d.). And this dataset is explained in detailed the features and target variables, But I choose the DIQ010 (Respondent is diabetic) because as I check the unique of this column it showed me the 0, 1, 2 numerical value which make me interested to used this as my feature and target variable.

As I begin to load the dataset using the df.head() code, It help me to show the first 5 rows and the 10 columns that help me to understand which column has the categorical variable and which columns has the numerical variables.

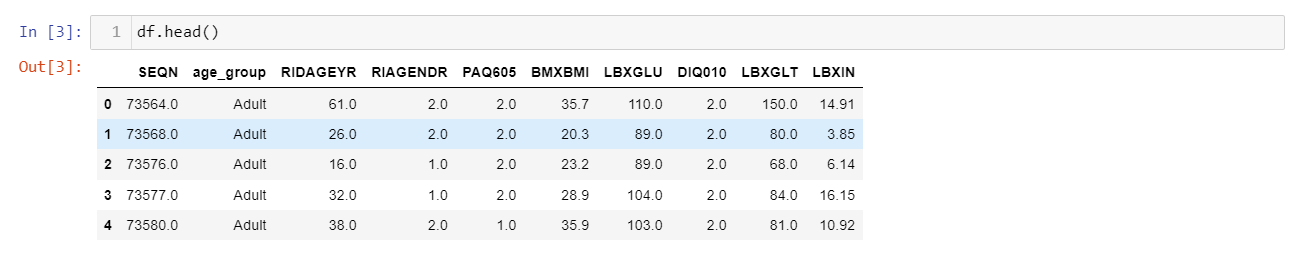


Figure 1: Head of the dataset

To fully understand the dataset here is the data dictionary that has the columns names, definition and the data types.

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| Columns name | Definition | Data Types |
| SEQN | Respondent Sequence Number. | Float64 |
| age\_group | Respondent's Age Group (senior/non-senior). | object |
| RIDAGEYR | Respondent's Age. | Float64 |
| RIAGENDR | Respondent's Gender. | Float64 |
| PAQ605 | If the respondent engages in moderate or vigorous-intensity sports, fitness, or recreational activities in the typical week. | Float64 |
| BMXBMI | Respondent's Body Mass Index. | float64 |
| LBXGLU | Respondent's Blood Glucose after fasting. | Float64 |
| DIQ010 | If the Respondent is diabetic. | float64 |
| LBXGLT | Respondent's Oral. | float64 |
| LBXIN | Respondent's Blood Insulin Levels | float64 |
| Reference | Reference for these columns and definition of the dataset (archive.ics.uci.edu, n.d.). | object |

In Figure 4 using the df.isnull().sum is very important to determine and evaluate if there is null values of the each columns and rows. The code df.isnull().sum() help me to understand that the dataset has no null values. In order for me to double check and make sure that completely has no null values I used the df.duplicated and it give me the results of false meaning there is no null values in the dataset that I want to do my data analysis.

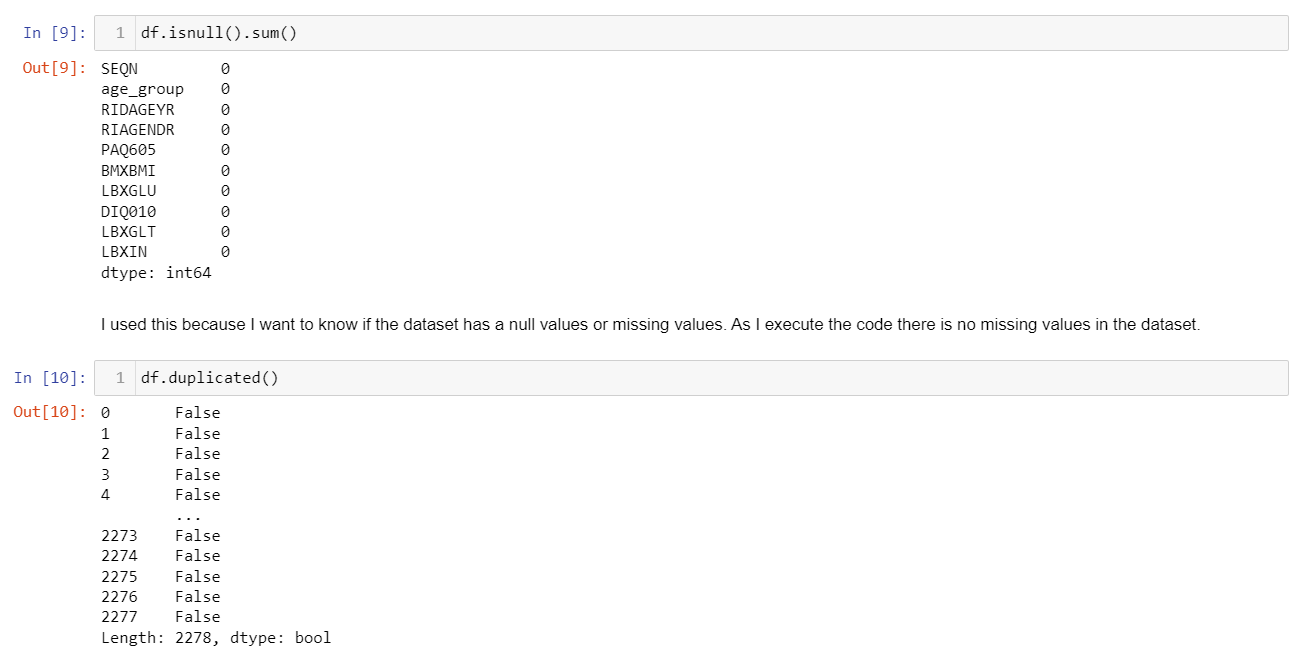


Figure 2: Using isnull().sum() and duplicate().

Proceeding to the next step which is the summary statistics of the data frame which is the df.describe(). It is responsible for the central tendency dispersion and the std shows the amount of change in the data and determines how it expands the values that come from the mean. The min shows the values of each column from higher to lower. Also, it helped me understand the numerical columns properly. (pandas.pydata.org, n.d.).



Figure 3: Describe of the dataset.

In figure 3, it is thorough show that the columns provide the numerical summary statistics of different columns and it will help me to better understand that if the value of the standard deviation of the columns is high meaning the data has a huge spread out, and if the standard deviation is low, it will be less spread out.

**Data Preparation**

The third phase is data preparation, which means that data cleaning should be implemented in this phase of the CRISP-DM methodology. It requires fixing the errors in the CO2 adsorption dataset before accepting the machine learning models. It has steps that need to be followed, like selecting data, cleaning the data (missing or outliers), constructing the data, integrating the data, and formatting the data. Some of the data has issues like outliers and missing values. (Saluja, 2018b).

It is vital to remove or drop the unnecessary columns that is not needed in further analysis of the dataset.

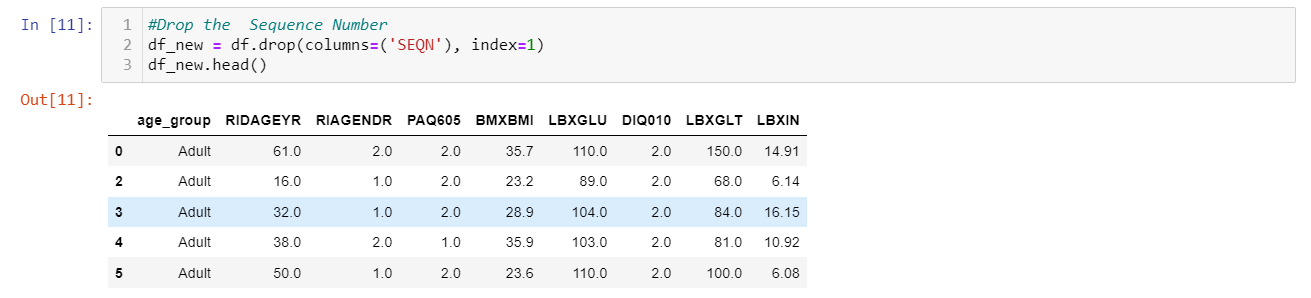


Figure 4: Dropping the column.

It is important to convert the categorical variables to numerical variables in preparation for the machine learning modelling. It is because machine learning only read numbers and process the analysis of accuracy score by reading only the numerical variables of each feature and target column.

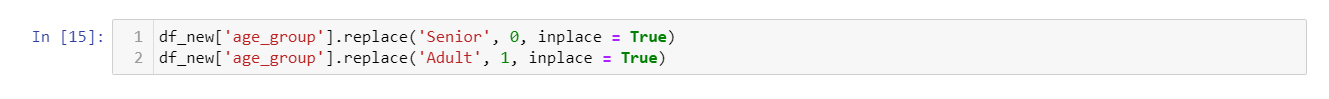


Figure 5: Converting the age\_group to numerical variable.

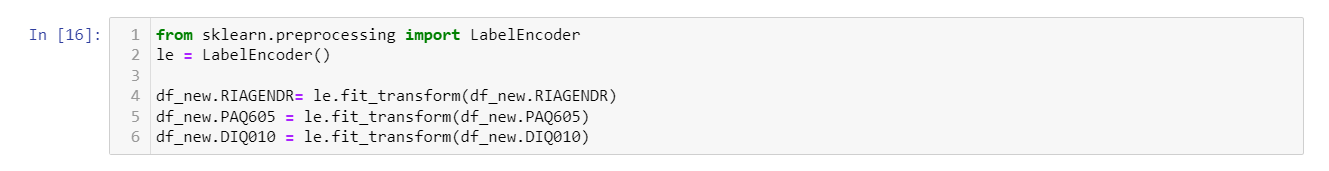


Figure 6: Label Encoding of the dataset.

In figure 7: I check the unique value of the column DIQ010 and it give me the results of 1, 0, 2 that makes me interested to used this as target and feature columns. As this give me the three numerical values it will be ideal to use as class features. In figure 8 as I used the label encoding it give me the constant value of features of 1.



Figure 7: checking the unique value of DIQ010 column of dataset.

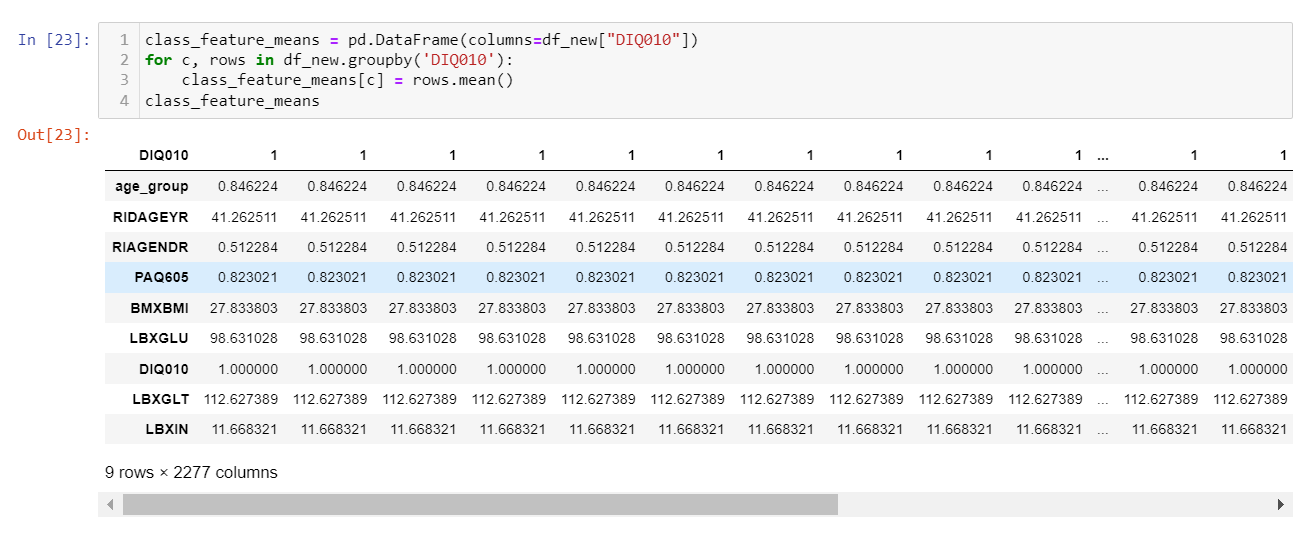


Figure 8: Class Features of the dataset.

It is vital to do the Linear discriminant analysis and principal component analysis to determine and compare which is the most suitable to used in machine learning modelling.

In figure 9 the linear discriminant analysis provide to showed from the 6 variance in y- axis LD2 , as I have seen the variance graph is going down and it is not good to use for machine learning modelling the LDA because the group is going down and as I analyse it will cause underfitting of the model.

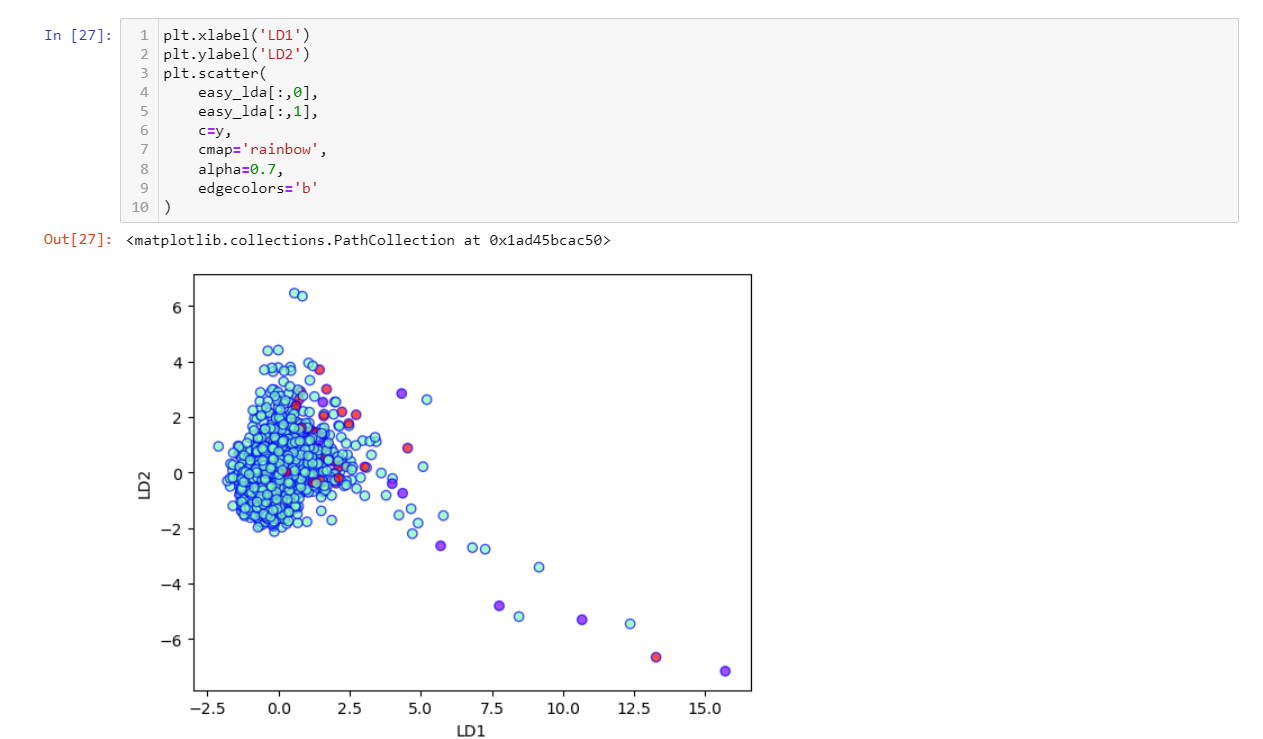


Figure 9: Linear discriminant analysis variance graph.

In Figure 10 Its is clearly showed that the variance graph is going up which means that this is the ideal to be used because it will give the efficient accuracy of the model. In case if will cause overfitting, I can use the GridSearch CV to reduced the overfitting of the model.

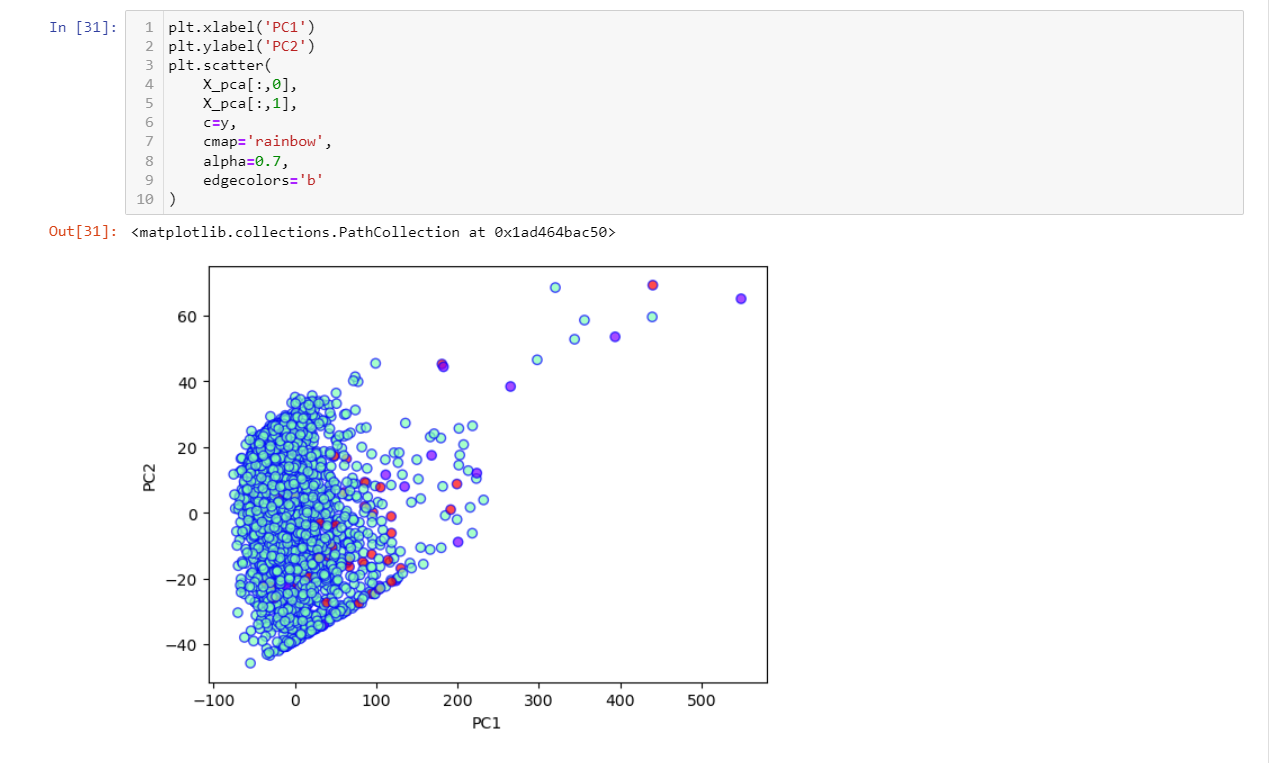


Figure 10: Principal component analysis variance graph.

**Modelling**

The first machine learning model that I used is the KNN, it is known for the simple to execute the classification algorithm. This can also used for the regression complication. This will calculate the distance for the input cognition with all observation during the training test.

In figure 11 it is clearly visualized in the confusion matrix there is false positive means that the respondent diabetic is test yes diabetic but in reality, they are not diabetic, and it has predicted values of 3. Also, there is false negative which means that the respondent diabetic is test not diabetic and in reality, it is not diabetic which has the predicted value of 2. Lastly, there is false negative that has the predicted value of 10 which means that the respondent diabetic is tested not diabetic and in reality, it is not diabetic.

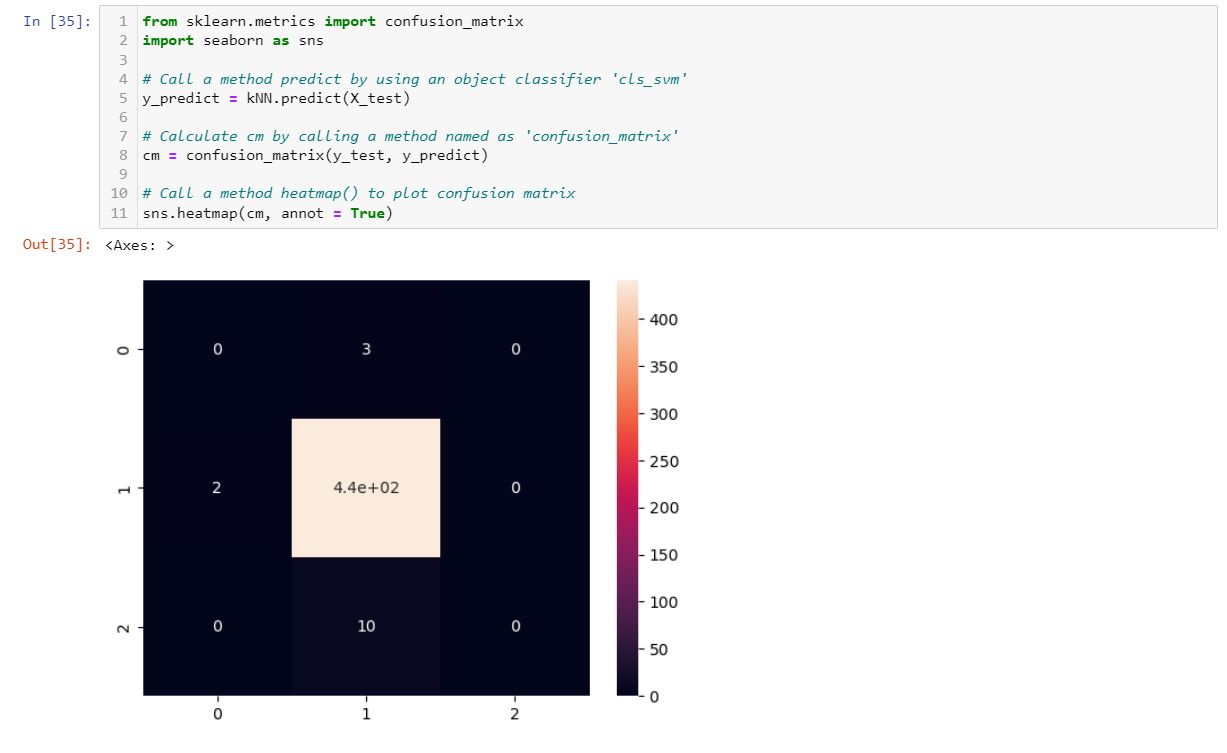


Figure 11: Confusion matrix of KNN classifier.

In figure 12 it is clearly stated on the graph that the accuracy f 1 score of the KNN classifier is 0.97, this is based on the varying number of neighbors. This is

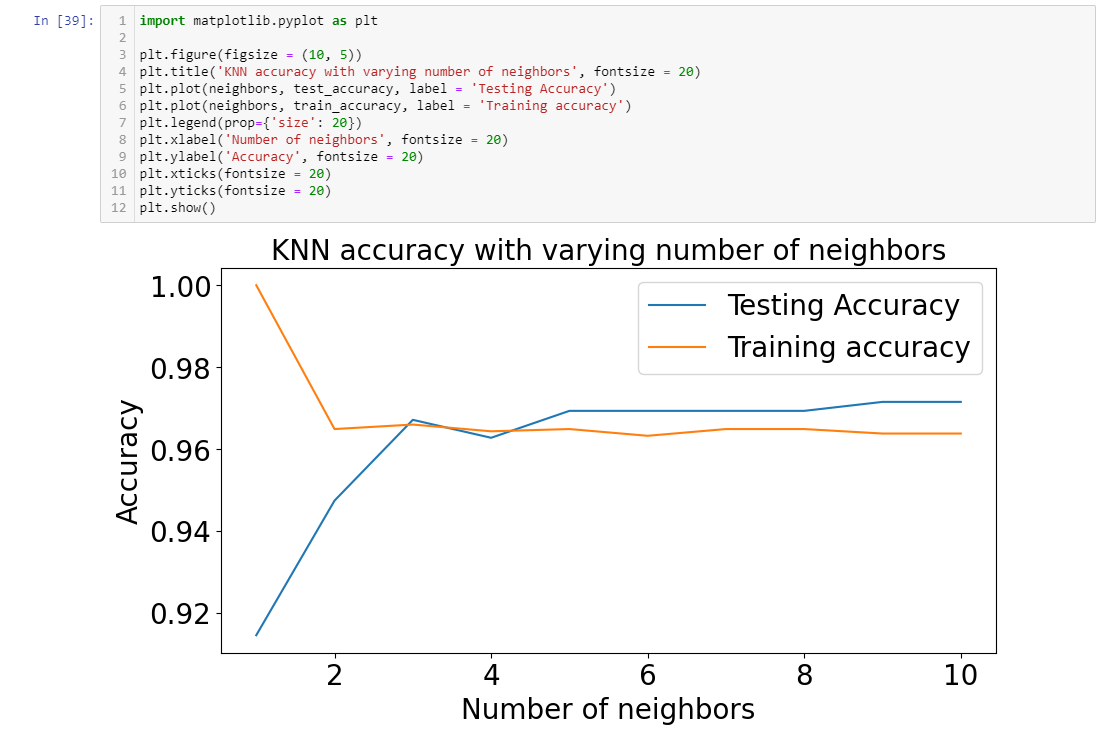


Figure 12: Accuracy graph score of KNN classifier.

In figure 12 I visualize the nearest neighbors by y-axis as target and x-axis as feature using 3 nearest neighbors on each test data. It shows that there are 5 nearest neighbor connected with black straight line into the test prediction at the center of each test data as it clearly stated in the plot.

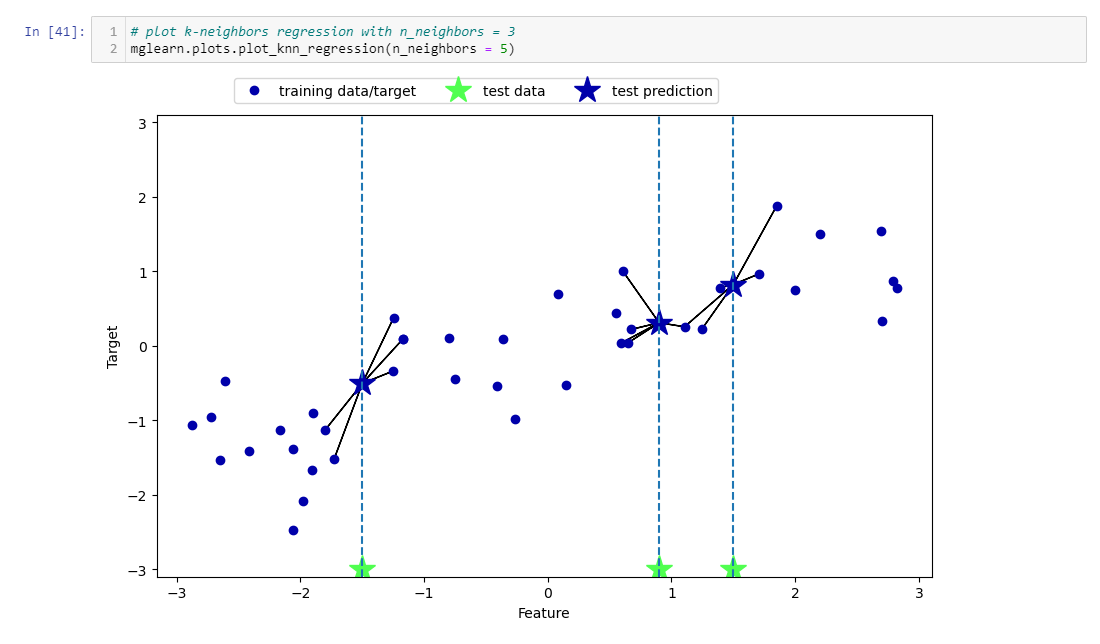


Figure 12 : K nearest neighbors regression.

The decision tree classifier is used to forecast the missing values, and it also has the ability to apprehend the non-linear pattern. It can also produce an overfit, and it is also easily affected by corrupt data. (Avinash Navlani, Fandango and Idris, 2021b). The Decision Tree Classifier is doesn’t need to do the standardize the features.

In figure 13 the decision tree classifier using the max depth of 3 has the f1 score accuracy results of 0.96. As I used the lower max depth it will give me the precise accuracy score of the decision tree classifier.

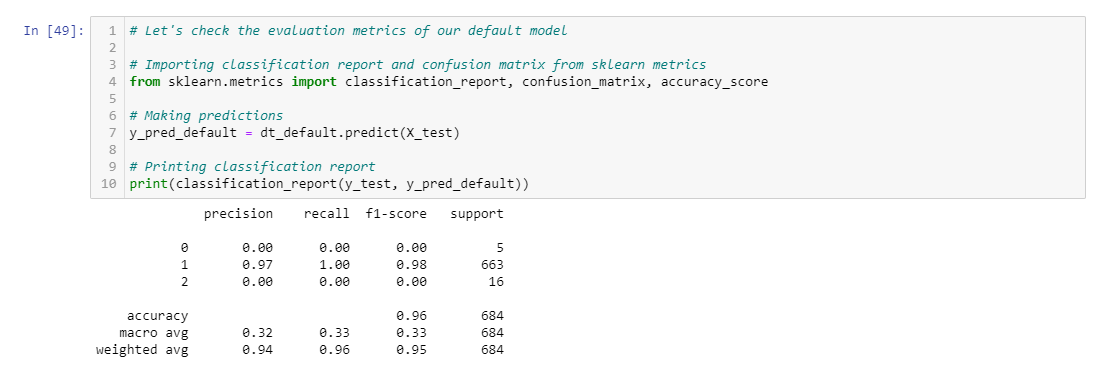


Figure 13: Decision tree classifier using max depth 3.

In figure 14: The decision tree classifier has the results of the multi class classification of the confusion matrix. It has the false positive which means that the respondent diabetic is test yes diabetic but in reality, it is not diabetic which has the predicted value of 5. In the first class second row of confusion matrix there is a false negative having a predicted value of 3 which means that the respondent diabetic is tested not diabetic but in reality, it is not diabetic. On the second class in second row the respondent diabetic is tested yes diabetic and in reality, it the result of yes diabetic which has the predicted value of 660. At the second rows of the second column, it has the false negative with a prediction value of 16 which means that the respondent diabetic has tested not diabetic but in reality, it is not diabetic.



Figure 14: Confusion matrix accuracy of decision tree classifier.

In Figure 15 A and 15 B This is vital to implement the GridSearchCV library in preparation to properly execute the reduction of the accuracy results using the max depth and criterion. As the max depth 9 is used this will provide a better reduction result of accuracy of the decision tree classifier which give me accuracy score of 0.93.

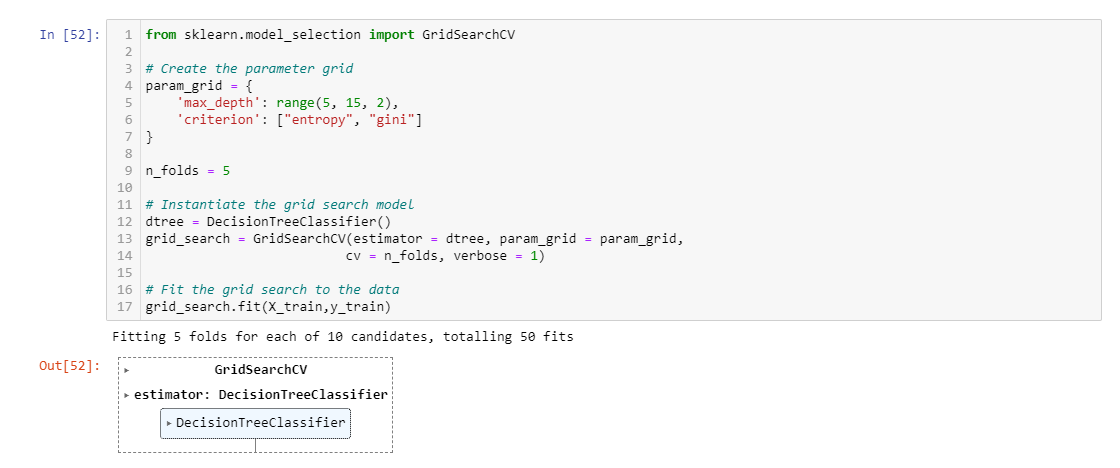


Figure 15 A: GridSearchCV of decision tree classifier.

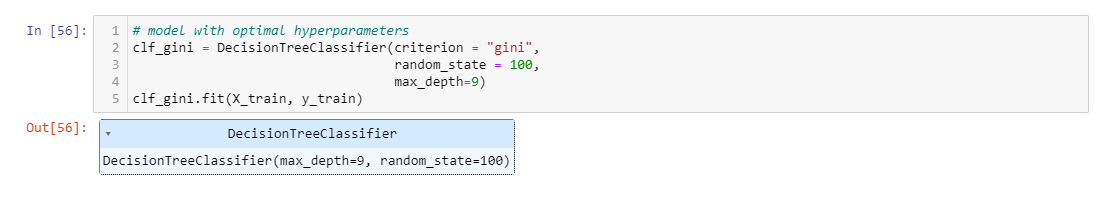


Figure 15 B: Hyperparameter using max depth of 9.



Figure 15 C: Accuracy results of using GridSearchCV max depth 9.

In figure 16 I have analysed and understand the by using the max depth of 9 this will result to plot the tree to have more branching of tree nodes and appears more compress and appear smaller as stated on the tree plot image.

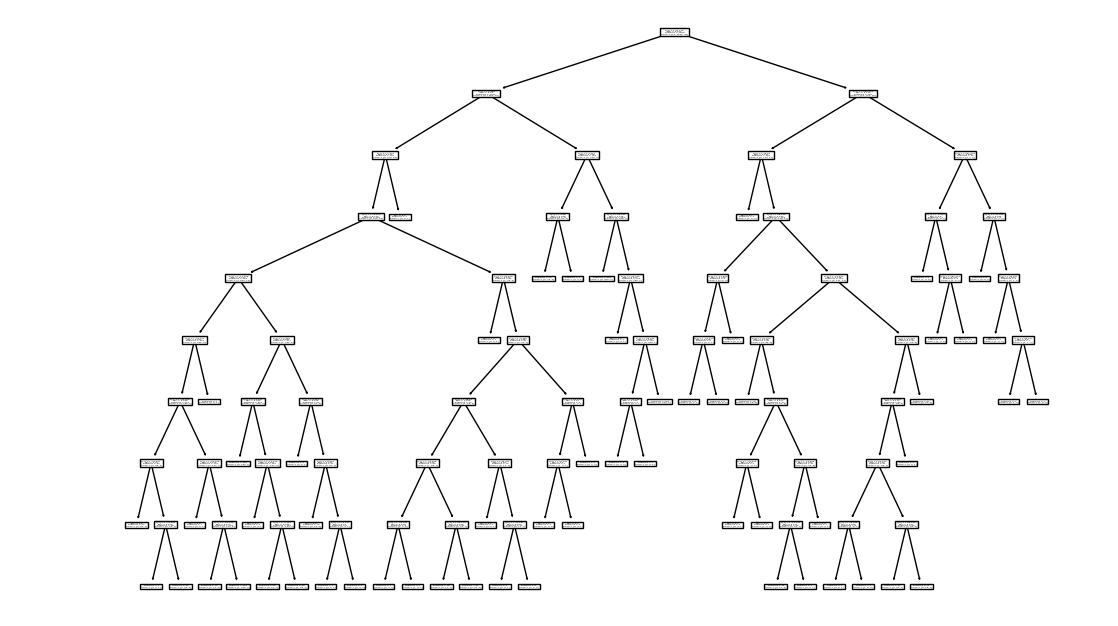


Figure 16: Decision tree classifier using max depth of 9 tree plot.

References List:

archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository*. [online] Available at: <https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)+age+prediction+subset>.

pandas.pydata.org. (n.d.). pandas.DataFrame.describe — pandas 1.0.3 documentation. [online] Available at: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html>.

Saluja, C. (2018b). *Data Preparation — A crucial step in Data Mining*. [online] Medium. Available at: https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281#:~:text=According%20to%20CRISP%2DDM%2C%20the..