*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 popular topics in the United States is the health and nutrition of the people. Since there are a lot of cases of unhealthy lifestyles of the food that they eat, there are many cases of a different age that has nutrition problems.

The obesity can lead to a serious disease in the United States, cases of obesity has increased the severe risk of having a health condition of high blood pressure, diabetes, stroke, and heart disease. There are many cases of type 2 diabetes in the United States because there are many people who eat unhealthy food that causes them to be overweight.

The report aims to have a thorough comparison of which of the LDA and PCA is more accurate to be used to train the model. It will be forecast by using KNN, Decision Tree Classifier, Random Forest, and K-means clustering.

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

Table of Contents

Business Understanding5

Data Understanding6

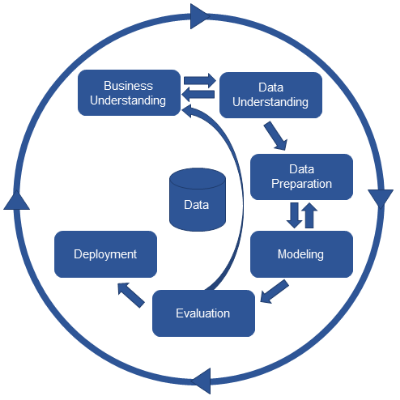
Data Preparation8

Modelling**11**

Evaluation20

Deployment20

References……………………………………………………………………21



**Business Understanding**

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

The principal objective of this project is to predict the respondent diabetic. This provides the comparison prediction of which of the LDA and PCA is most accurate to be used in the preparation of the machine learning model. The hypothesis of this project is to compare the accuracy results of the machine learning models to determine which one provides the most sufficient 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 overfitting and underfitting of the models.
* Implement the chosen machine learning model to predict the respondent's diabetes using the PCA to have a better accuracy score.

General Goal:

 The project aim to asses some machine learning models to predict the respondent diabetic in class 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

* Modeling: KNN, Decision Tree Classifier, Random Forest, and K-means clustering.

**Data Understanding**

 Data understanding is the second phase of the CRISP-DM Methodology, meaning what I have understood to the data.

In this second stage of the CRISP-DM, it is crucial to take some time to look at every detail dataset. To avoid encountering some errors when I will proceed to the data preparation. The Data preparation is the vital part of the data analysis which is the third phase of the CRISP-DM methodology. On that part, I need to execute the data analysis like cleaning the data which is vital before performing the modeling.

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

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

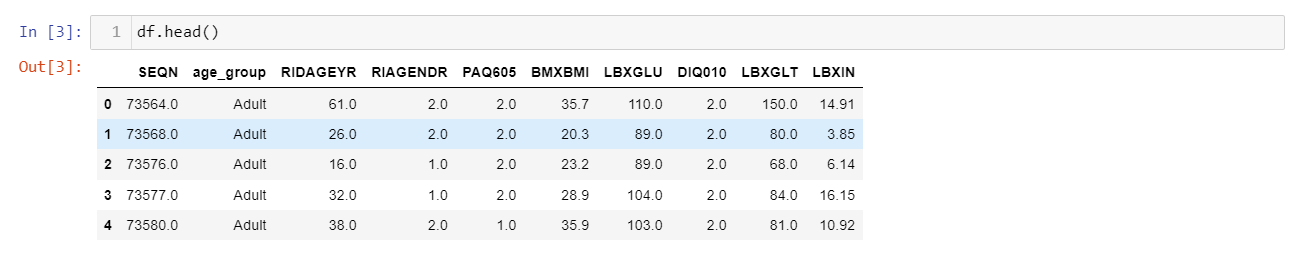


Figure 1: Head of the dataset

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 |

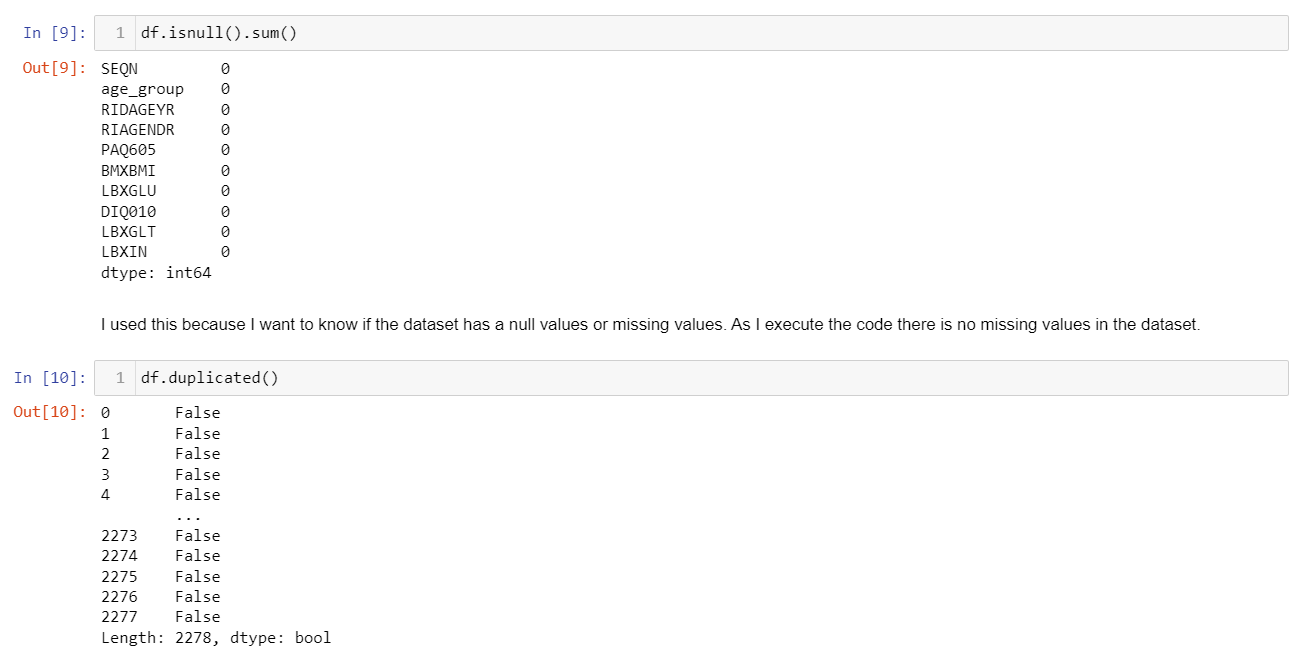
Using the df.isnull().sum is vital in determining and evaluating if there are null values of in each column and row. The code df.isnull().sum() helps me to understand that the dataset has no null values. For me to double check and make sure that there are no null values I used the df.duplicated it gave me the results of false which means there are 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 the summary statistics 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.

It thoroughly shows that the columns provide the numerical summary statistics of different columns and it will help me to better understand if the value of the standard deviation of the columns is high meaning the data has a high spread out, 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 column that is not needed in further analysis of the dataset.

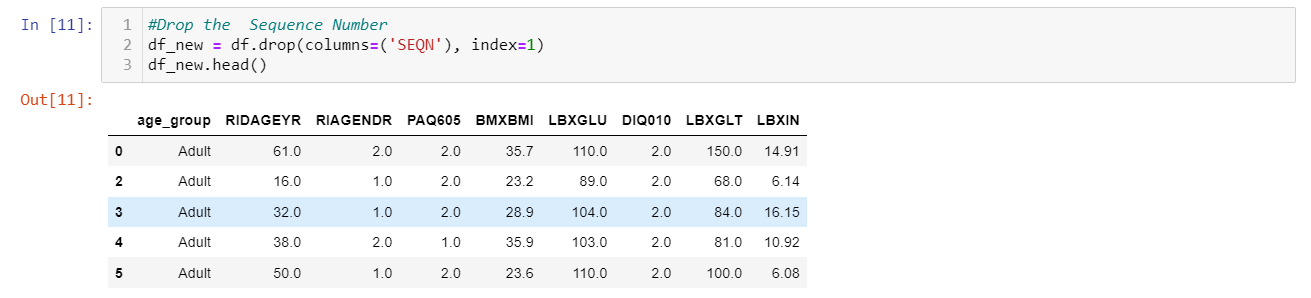


Figure 4: Dropping the column.

It is important to convert the categorical variables to numerical variables in preparation for the machine learning modelling. This is because machine learning only reads numbers and processes 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.

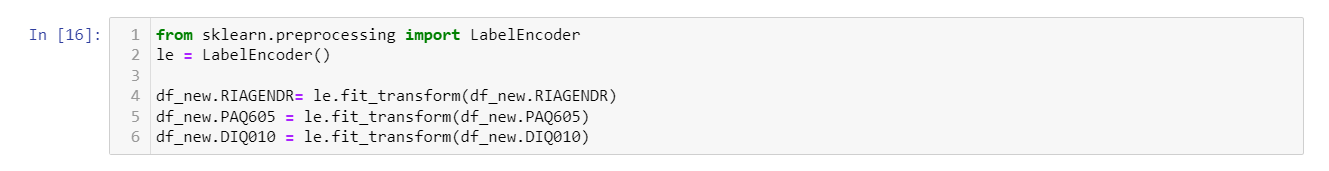


Figure 6: Label Encoding of the dataset.

I checked the unique value of the column DIQ010 and it gave me the results of 1, 0, and 2 which makes me interested in using this as target and feature columns. As this gives me the three numerical values it will be ideal to use as class features. I used the label encoding give me the constant value of features of 1 (Figure 7).



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

As I have analysed, it is important to know the relationship between the target variable in the numerical column and the categorical column, which is age group. In the age group, the senior has the least diabetic respondent, and the adult has the highest diabetic respondent. The RIAGENDR (respondent's age) has a lower value at 0 and a higher variance in the bar graph at a value of 1. The PAQ605 has the least respondent diabetic in the value of 0 since these are the respondents that exercise every week, while the 1 value has a high variance graph because the respondent does moderate exercise every week.



Figure 8: Subplot of the numerical and categorical columns with the target variable.

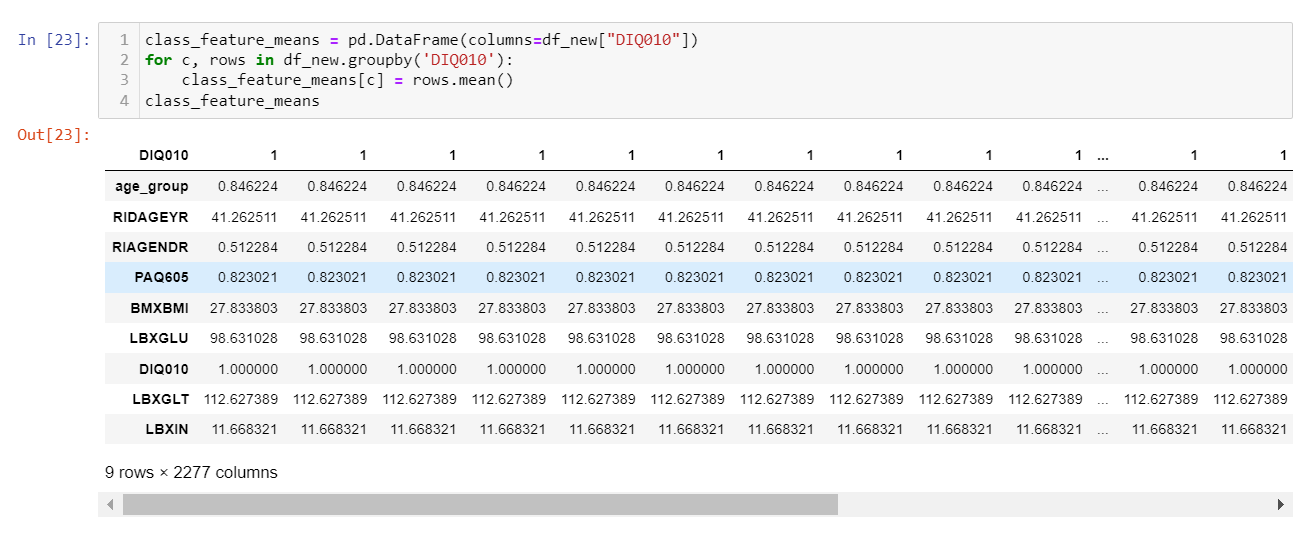


Figure 9: Class Features of the dataset.

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

The linear discriminant analysis provides to show from the 6 variances in y- the 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 class is going down and as I analyse it will cause underfitting of the (Figure 9).

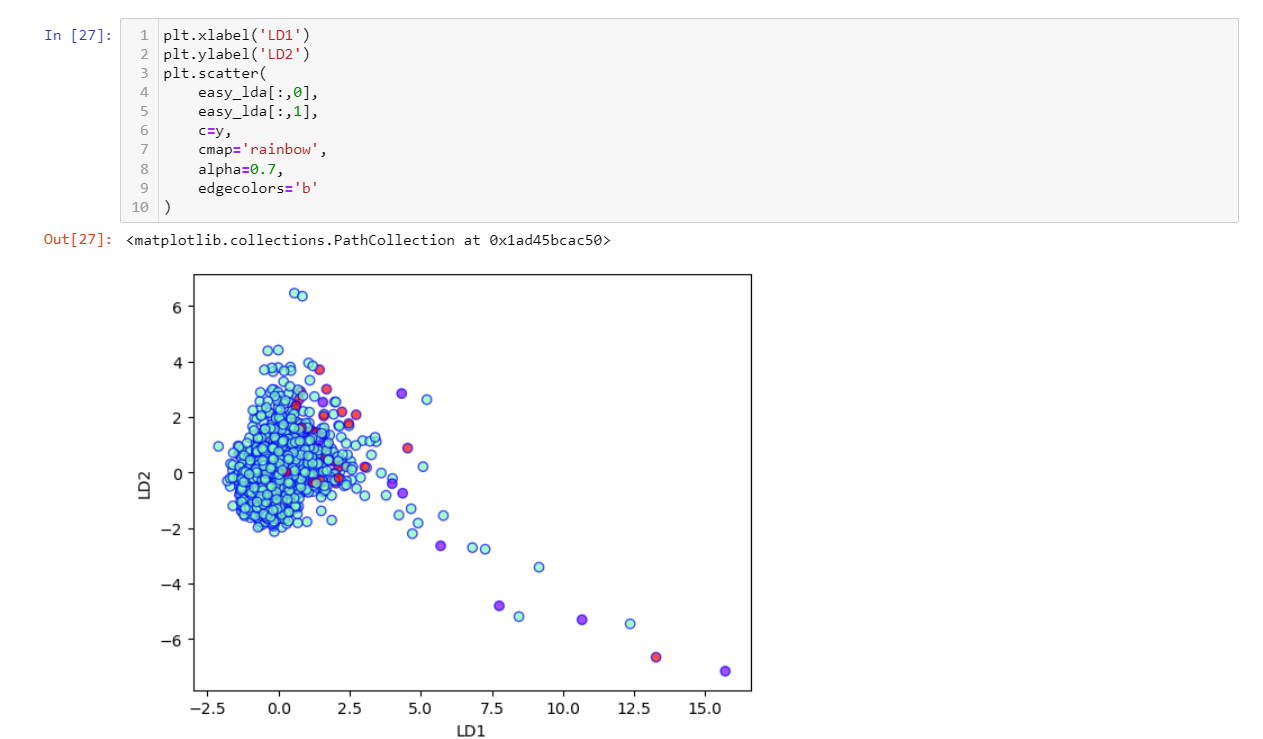


Figure 9: Linear discriminant analysis variance graph.

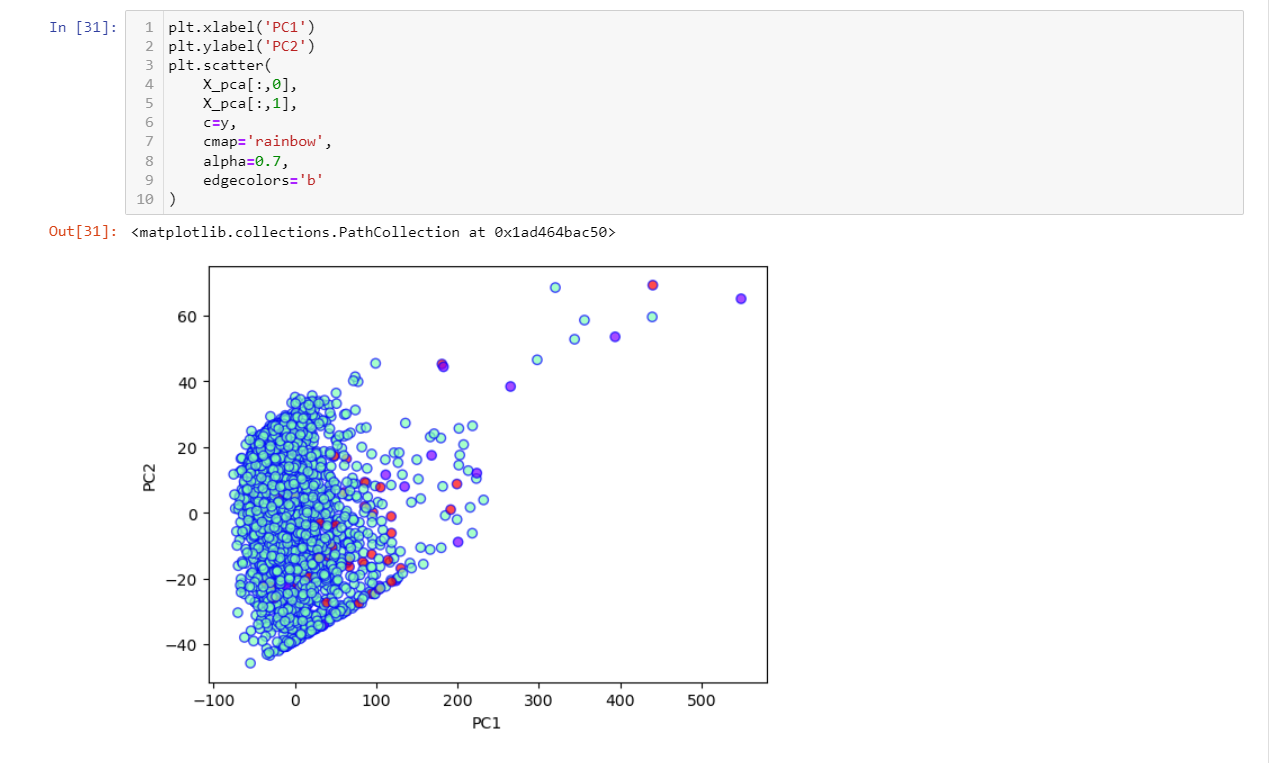
It shown that the variance graph is going up which means that this is the ideal to be used for machine learning model because it will give the good accuracy score of the model. In case it will cause overfitting, I can use the GridSearch CV to reduce the overfitting of the model (Figure 10). 

Figure 10: Principal component analysis variance graph.

**Modelling**

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

It visualized in the confusion matrix there is a false positive that the respondent's diabetic is test yes diabetic and in reality, it is not diabetic, and it has predicted values of 3. Also, there is a false negative which means that the respondent's diabetic is test not diabetic, and in reality, it is yes diabetic has the predicted value of 2. Lastly, there is a false negative that has the predicted value of 10 meaning the respondent diabetic is tested not diabetic, and in reality, it is yes diabetic (Figure 11).

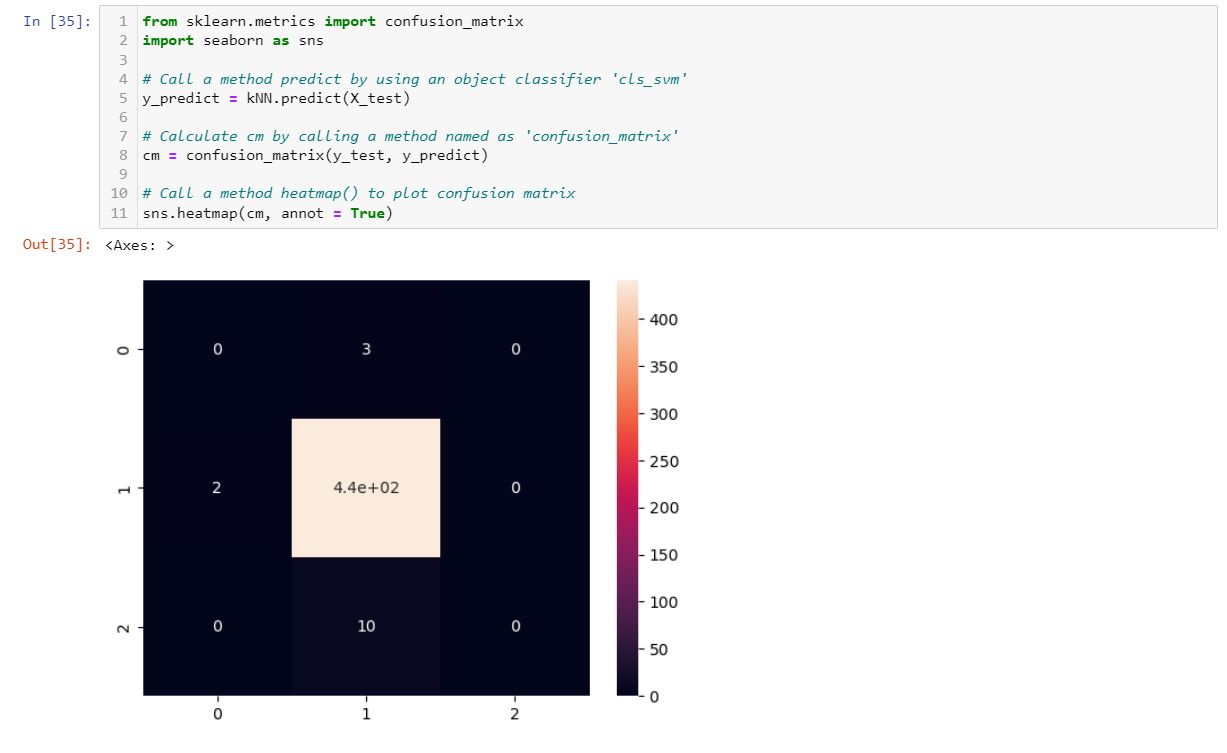


Figure 11: Confusion matrix of KNN classifier.

As stated on the graph the accuracy f 1 score of the KNN classifier is 0.97, this is based on the varying number of neighbors (Figure 12).

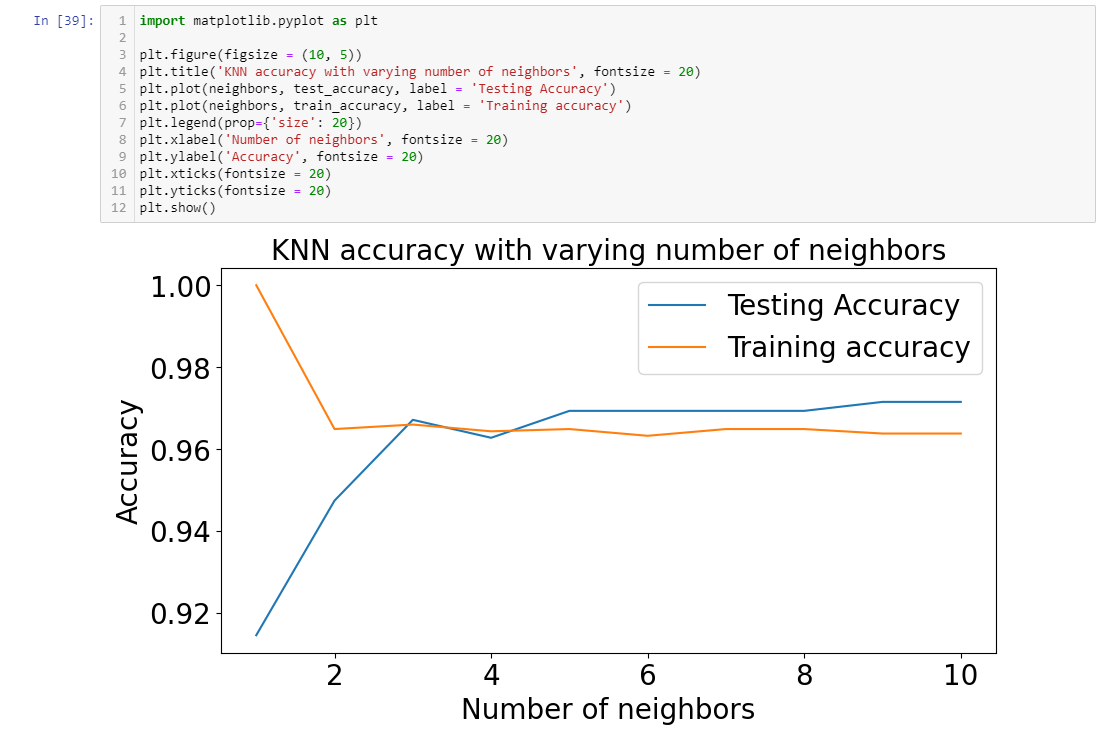


Figure 12: Accuracy graph score of KNN classifier.

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.

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.

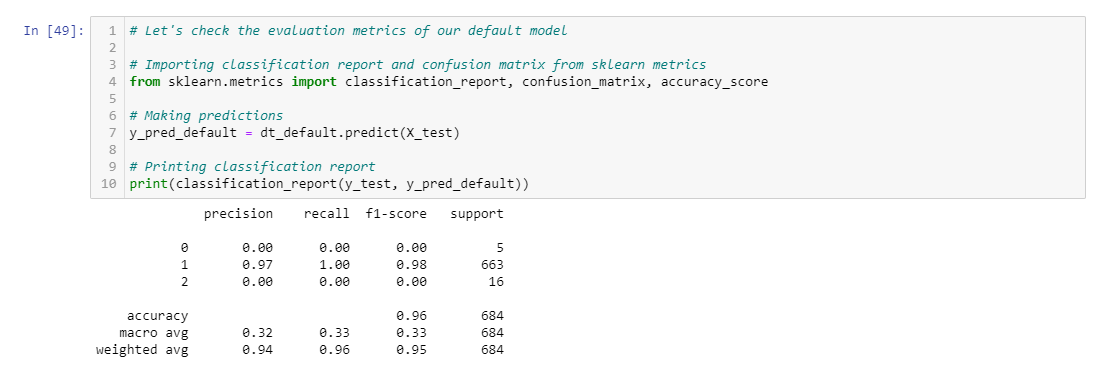


Figure 13: Decision tree classifier using max depth 3.

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



Figure 14: Confusion matrix accuracy of decision tree classifier.

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 the accuracy of the decision tree classifier which gives me accuracy score of 0.93.

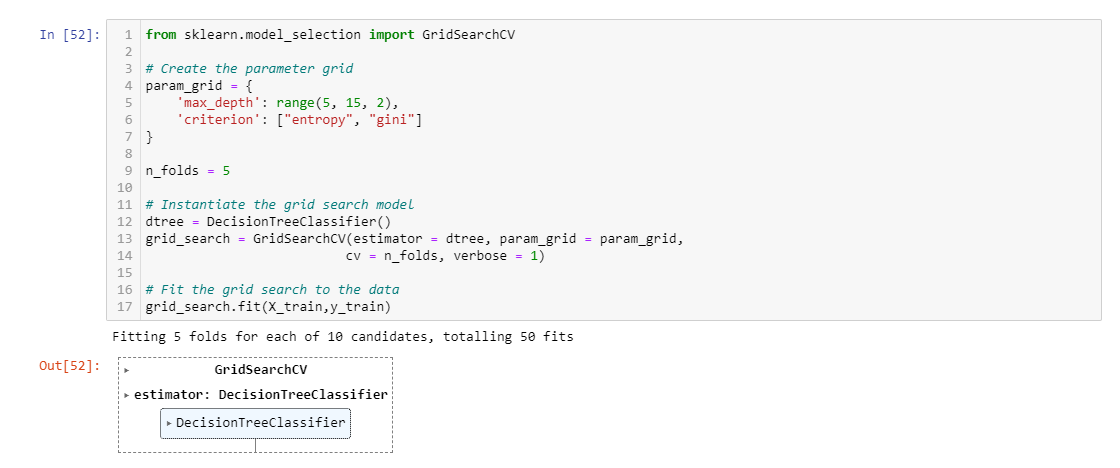


Figure 15 A: GridSearchCV of decision tree classifier.

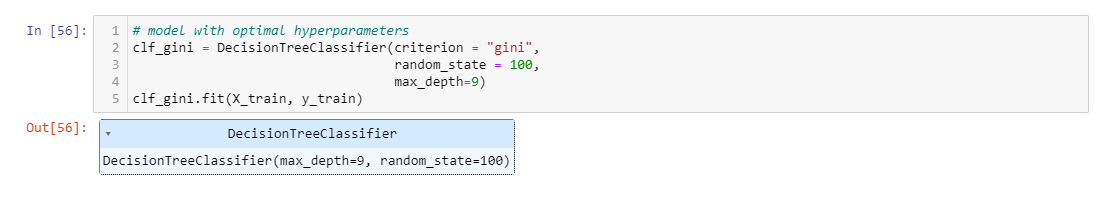


Figure 15 B: Hyperparameter using max depth of 9.



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

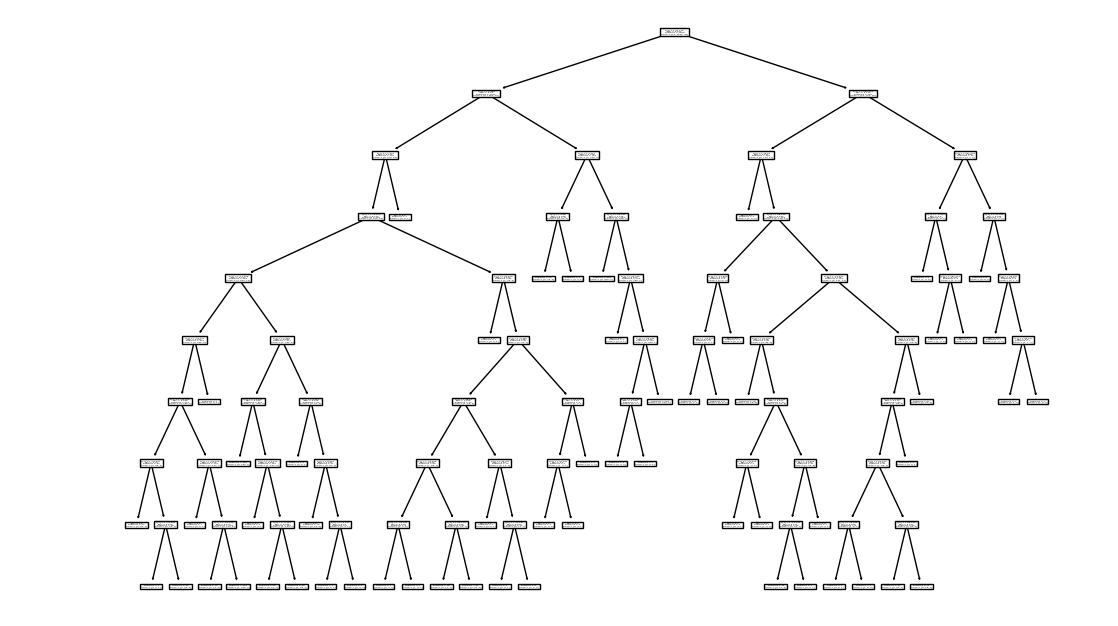
I have analysed and understand that using the max depth of 9 will result in the plot of the tree having more branching of tree nodes and appearing more compressed and smaller, as stated on the tree plot image (Figure 16).

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

The random forest classifier has a f-1 score accuracy of 0.96; this can be improved by using the GridSearchCV, which has a Kfold of 5. Using the max depth of 2, 40, and 5, there is a false positive respondent who is diabetic, yes, but in reality, it is not diabetic and has the value 6. In the class three column row one, the false negative respondent diabetic is tested as not diabetic, but in reality, it is diabetic, which has a value of 1. In class one, row two, the respondent diabetic has a false negative and was tested as not diabetic, but in reality, it is yes diabetic, which has a value of 1. In class three, row two, the respondent diabetic has a false positive, which means that tested yes is diabetic, but in reality, it is not diabetic, which has a value of 8. In class two, row two, the respondent is diabetic, which has a true positive meaning that it was tested as yes, diabetic, and in reality, it is diabetic. Last but not least, in class two, row two, the respondent's diabetes is false negative, which means he was tested as yes diabetic, but in reality, he is not diabetic (Figure 17 B).

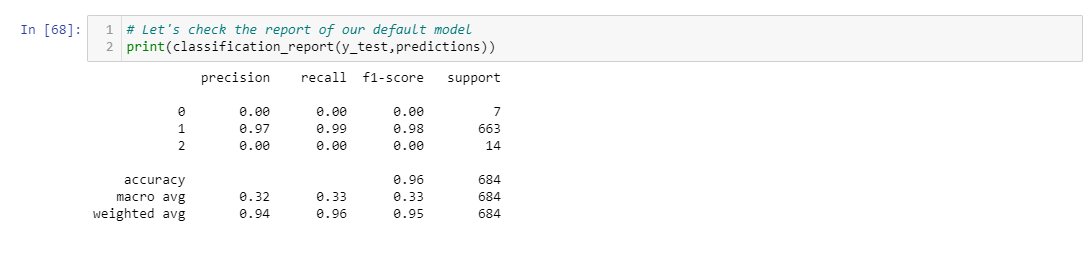


Figure 17 A: Accuracy score of the Random Forest Classifier.

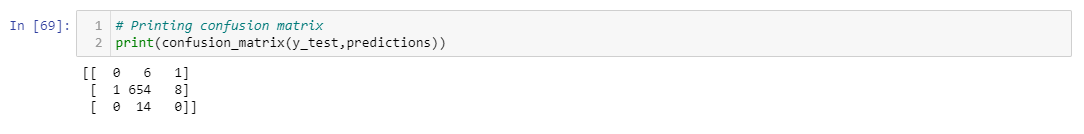


Figure 17 B: Confusion matrix of the random forest classifier.



Figure 17 C: GridSearchCV of the random forest classifier.

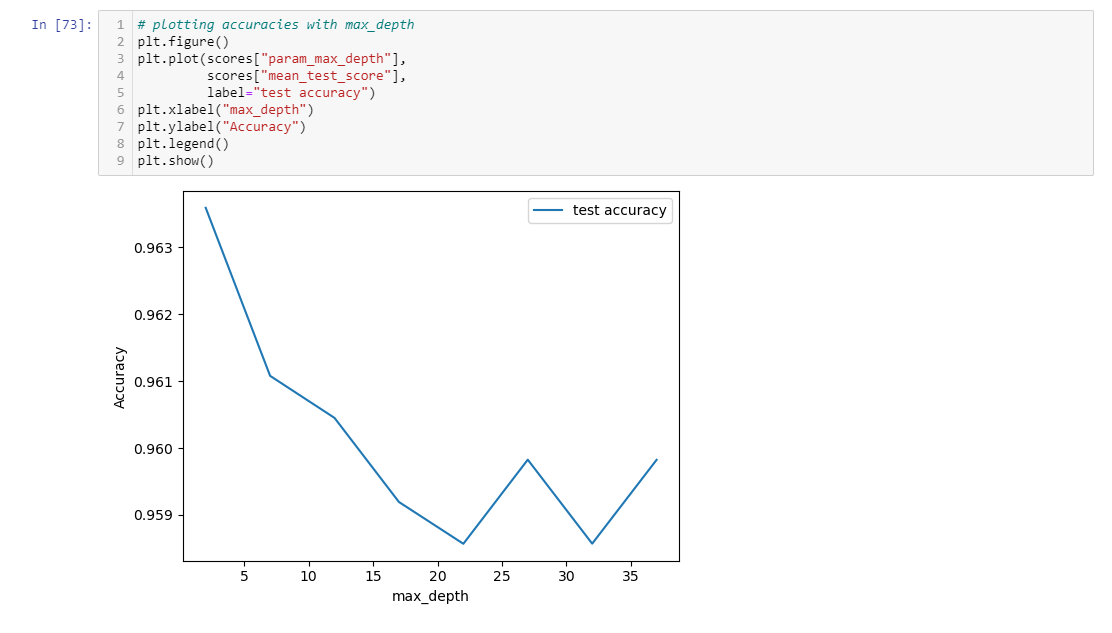
Using the higher max depth, it will give me a graph like an elbow graph, and as I increase the max depth, it will reduce the accuracy value from 0.97 to 0.95, as stated on the test accuracy graph (Figure 17D).

Figure 17 D: Accuracy plot of the max depth.

The kmeans clustering has the prediction of the three clusters. As I analyse and understand, cluster 0 has the most accurate kmeans because the points of the Euclidean distance are well defined and accurately visualised than clusters 1 and 2. In cluster 2, it is shown that the plots are scattered, while in cluster 1, they are intact and a little scattered on the left side (Figure 18).

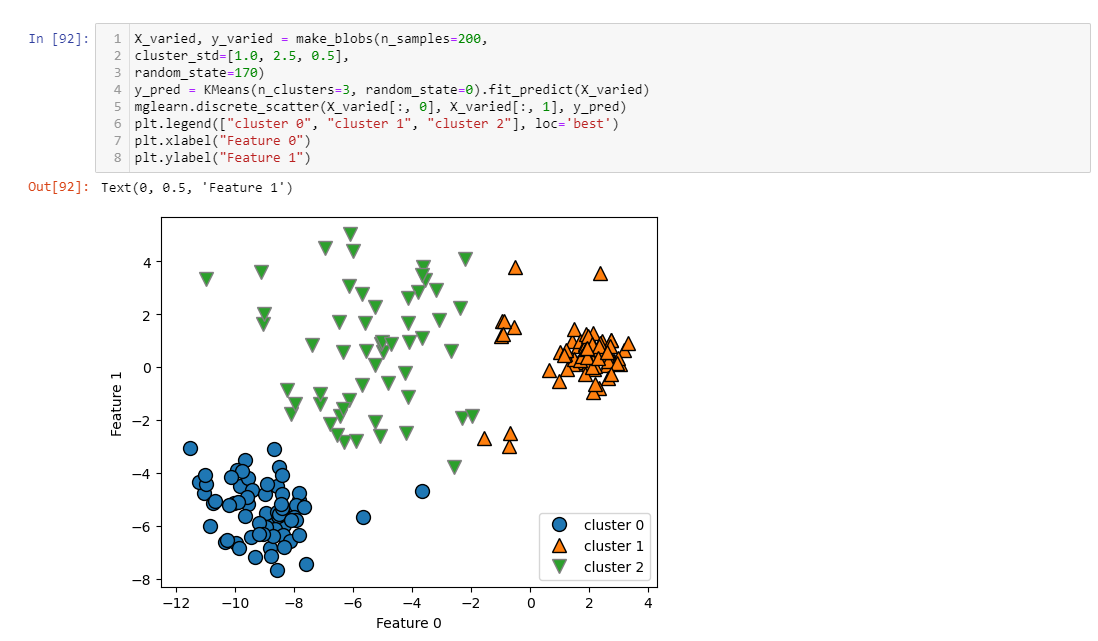


Figure 18: K means clustering plot graph.

I just want to experiment with the age group and the LBXGLU (blood glucose after fasting) to know which age group has the highest rate of glucose after fasting, which will be determined in the plot. As shown in the plot graph, the age group has a lower variance with the senior age, which has a 0 value on the x-axis, than the adult age, which has 250 variances on the y-axis. While the adult, which has a value of 1 on the x-axis, has high blood glucose after fasting because it has a variance of 400 on the y-axis (Figure 19).



Figure 19: Experiment of the age\_group and LBXGLU (Blood Glucose after fasting) columns.

  It is stated the comparison of the accuracy score of the silhouetter score using different kmeans nearest cluster. This helped me understand that, as I used the 2 kmeans nearest cluster, it would give me a higher silhouetter score of 0.538. Also, as I used the kmeans nearest cluster 3, it will give me a moderate silhouetter score of 0.416. And lastly, if I use the kmeans nearest cluster 4, it will give me a lower Silhoutter score of 0.314. The higher the kmeans nearest cluster, the lower will be the results of the silhouetter score (Figures 20A and 20B).



Figure 20 A: Silhouetter score first two results.



Figure 20: B Silhouetter score three results.

It is well determined within cluster sum of square, which is on the y axis. As shown in the graph, cluster 1 within the cluster sum of squares is very high in the 7-variance value on the y axis. Then, as it moves to the second cluster, it reduces the variance to 4 on the y axis. As it moves to the third cluster, it shows that it has a reduced variance of 3 on the y axis. Until it moves to a larger cluster, it will gradually reduce the within-cluster sum of squares value on the y axis. This is perfectly called the elbow method (Figure 21).

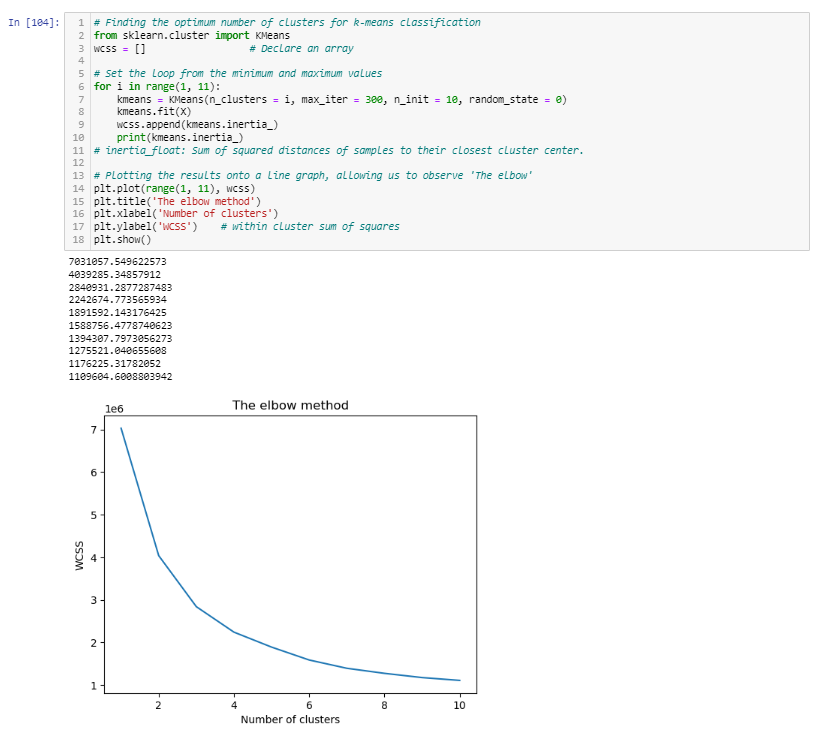


Figure 21: Finding the optimum number of clusters for k-means classification.

**Evaluation**

Since the accuracy score is compared to different models to know which models will give sufficient accuracy using the GridSearchCV, we can reduce and improve the model accuracy score. The results show that the hypothesis is correct enough to improve the models using the decision tree classifier with a max depth of 9 that gives accuracy results of 0.93, which will provide safety in overfitting and underfitting of the models.

Therefore, I conclude that the dataset for this CA2 has challenged me to do my data analysis. Even though I encounter some errors, I try my best to fix it. I can say I really learned a lot from this CA2 that I can use and apply for my next assignment in the future.

The next steps for this project will be:

* Experiment with columns that I can set as target variables.
* Apply another supervised machine learning method to train the model.

**Deployment**

Reviewing this CA2 assignment, I can definitely say it is harder than I was thinking and expecting. It is because this CA2 is combined with two subjects, and I need to do my analysis in comparing the PCA and LDA to determine which of them is most suitable to use in training the models. No matter how hard I have experienced, I am still happy that I did my very best to finish the data analysis of my CA2.

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>.

Avinash Navlani, Fandango, A. and Idris, I. (2021). *Python Data Analysis*. Packt Publishing Ltd.

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..