**Prediction of Presence of Diabetes**

***Name: Kartik Ningaraj Yanigar***

***E-mail:kartik.nyanigar.mec20@itbhu.ac.in***

***Ph No: 8867021358***

***Project for Internship at Exposys Data Labs***

ABSTRACT

A machine learning model is a computer program trained to recognize specific patterns. You teach a model on a set of data and give it an algorithm to use to reason about and learn from that data.

Once the model has been trained, you can use it to reason over data it hasn't seen before and make predictions about it. Here this concept has been used to predict the profits of a company based on some already given features.

Table of Contents

1. Introduction
2. Existing method
3. Proposed Method with Architecture
4. Methodology
5. Implementation
6. Conclusion

Introduction

Diabetes is a type of chronic disease which is more common among the people of all age groups. Predicting this disease at an early stage can help a person to take the necessary precautions and change his/her lifestyle according to either prevent the occurrence of this disease or control the disease.

In this project we will be using the dataset called **“Pima Indians Diabetes Database”** by UCI Machine Learning Repository. This dataset is originally from the **National Institute of Diabetes and Digestive and Kidney Diseases.** The dataset has been extracted from [**kaggle.com**](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database) and the dataset has been uploaded by **UCI Machine Learning Repository.**

The dataset contains given features-

1. **Pregnancies** - showing the number of pregnancies a person had had.
2. **Glucose** - Glucose level of the person.
3. **Blood** **Pressure**
4. **Skin** **Thickness**
5. **Insulin** – Insulin level of person’s blood.
6. **BMI** – Body Mass Index of the person.
7. **Diabetes Pedigree Function** – scores of likelihoods of diabetes based on family history given Diabetes Pedigree Function.
8. **Age**
9. **Outcome** – A binary value (1 or 0) which indicates presence or absence of diabetes.

Existing Method

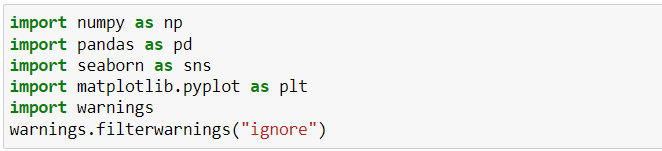
Existing dataset is a dataset with 8 independent variables namely [Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age] and an dependent variable Outcome (1 or 0) indicating the presence or absence of Diabetes.

Many Classification algorithms can be employed to predict the outcome like Logistic Regression, Support Vector Classifiers, Decision Tree Classifiers,

Logistic Regression and Support Vector Classification are widely used.

**Exploring Data**

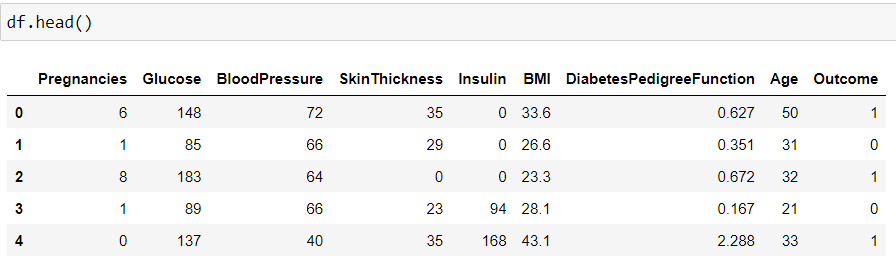
Firstly, importing required libraries to perform basic operations.

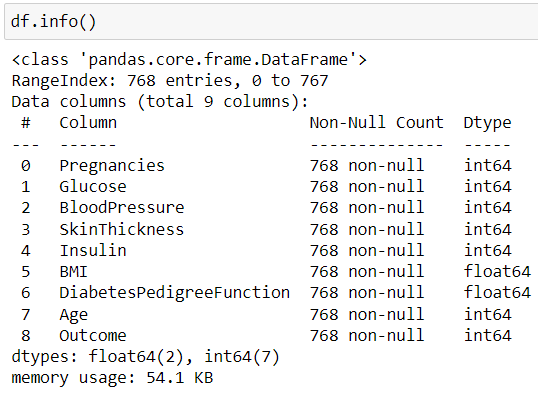


Reading the dataset into pandas’ data frame.



Thus, dataset looks like,



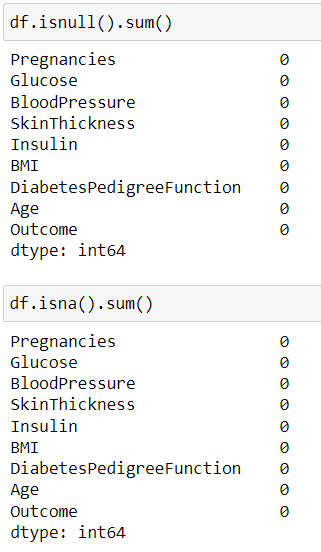


The figure above shows that there are no string or special data types. Thus, only integer and float data types to deal with.

**Cleaning Data**

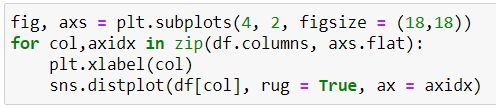
We need to consider several factors during data cleaning like duplicate observations, bad labelling, missing or null data points, outliers.

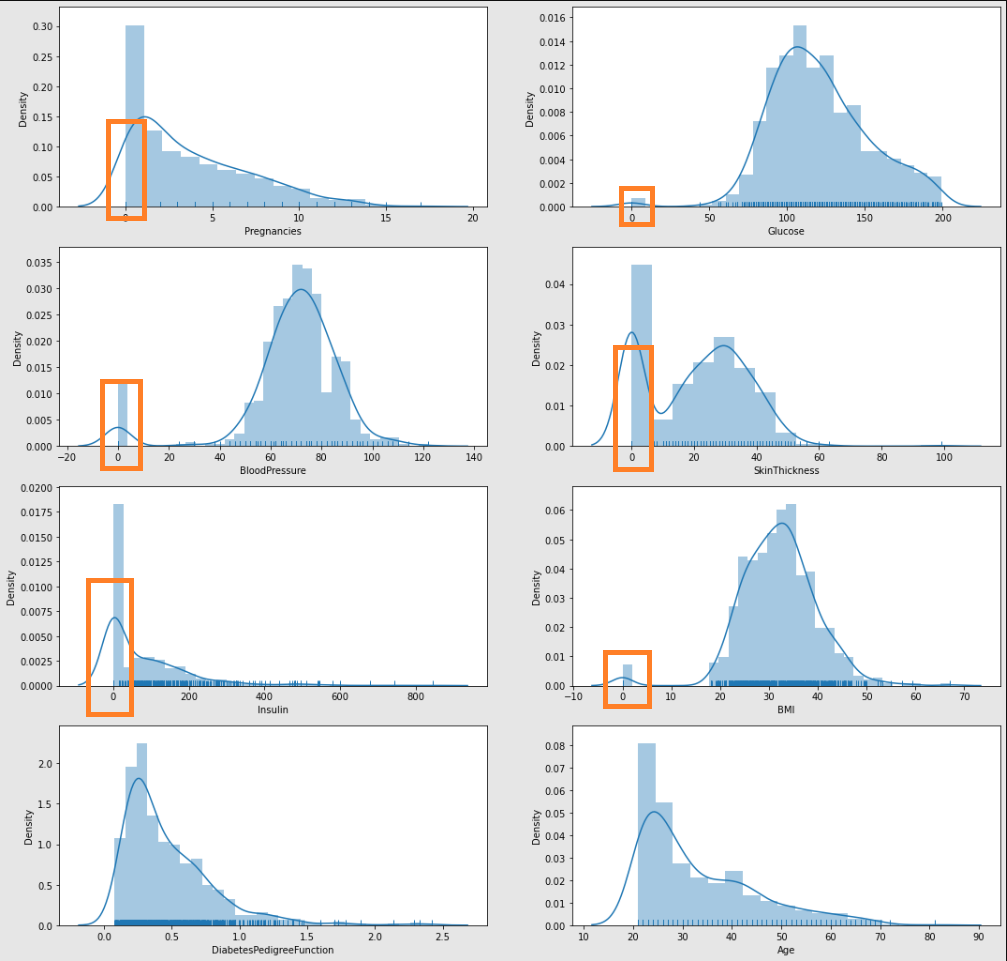
* Checking for null values in dataset.



Thus, above figure shows that the dataset doesn’t contain any missing or null values.

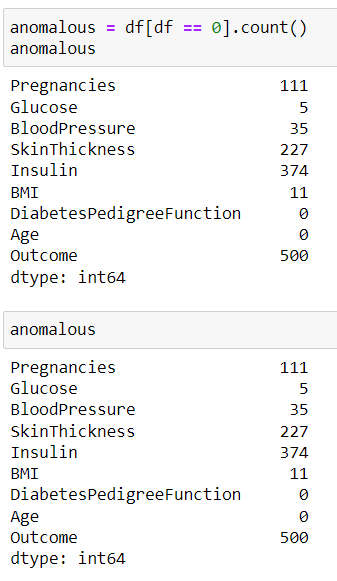
* Checking for Outliers in dataset

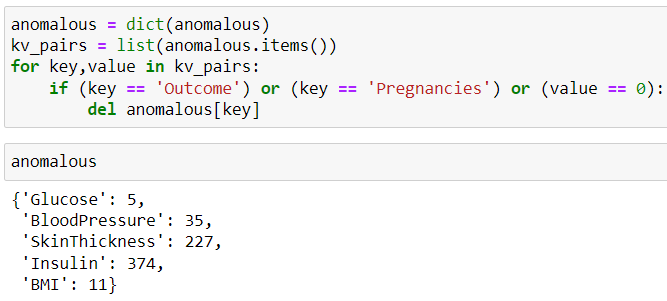




In the above figure the outliers are being indicated within the orange rectangles.

We are checking for number of outliers.

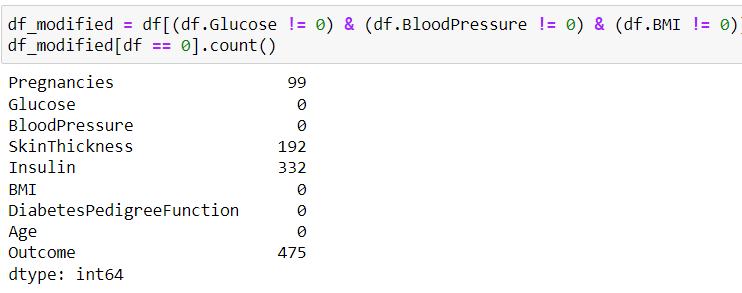




Thus, in above figure the few features have values as zeros whereas it impossible for **Glucose** levels to be zero, **Blood Pressure** level to be zero, **Skin** **Thickness** to be zero, **Insulin** to drop to zero, **BMI** of any person to be zero.

If any features had actually reached a zero level in these features, he or she would surely be dead.

Trying and removing the outliers that have little to no effect on dataset.



But still **Skin Thickness** and **Insulin** have zero levels and are not good features to be considered for classification. Thus, we eventually have to perform dimensionality reduction to access best features for classification.

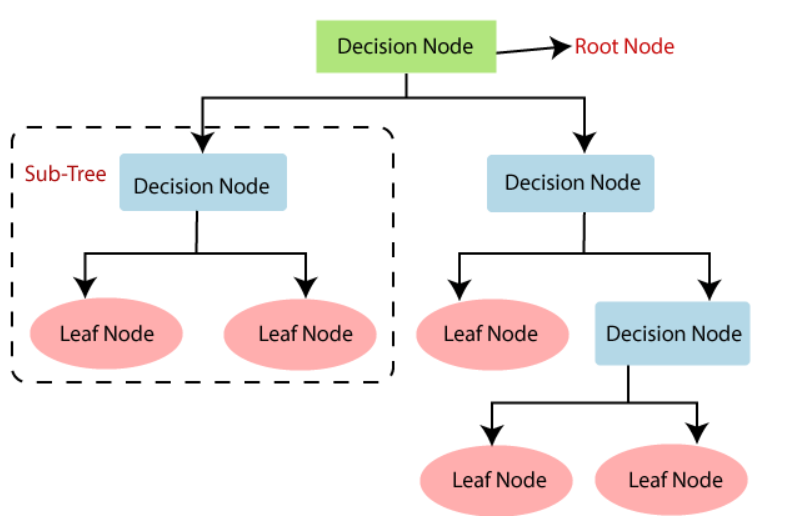
Proposed method with Architecture

Thus, according to industrial experts Logistic Regression and Support Vector Classification etc. is recommended. But the type of algorithms used depends on the dataset.

In this project Decision Tree Classifier is proposed.

**Decision Tree** is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**

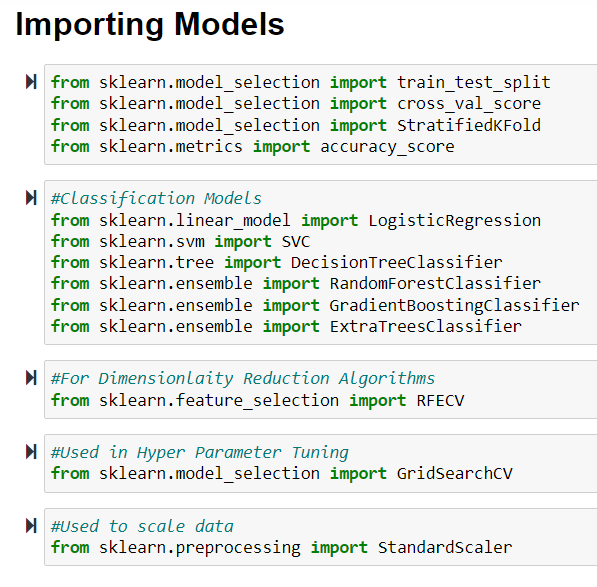
And Decision Trees can be visualised as:

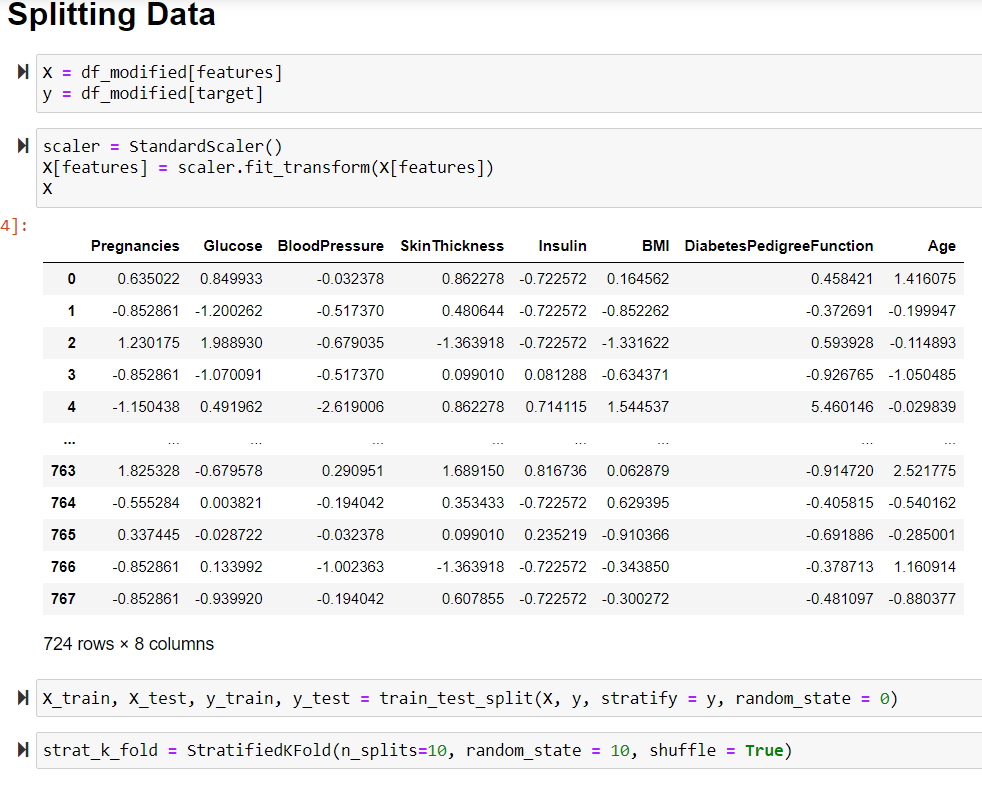


Methodology

1. Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
2. Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
3. **Decision Tree** is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
4. ***Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.***  Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
5. **Gradient Boosting Machine (GBM)** is one of the most popular forward learning ensemble methods in machine learning. It is a powerful technique for building predictive models for regression and classification tasks. GBM helps us to get a predictive model in form of an ensemble of weak prediction models such as decision trees. Whenever a decision tree performs as a weak learner then the resulting algorithm is called gradient-boosted trees.
6. **Extremely Randomized Trees Classifier (Extra Trees Classifier)** is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Implementation



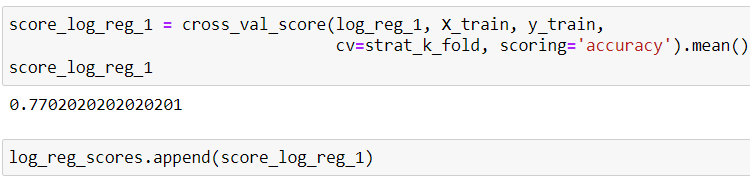


1. **Logistic Regression-**

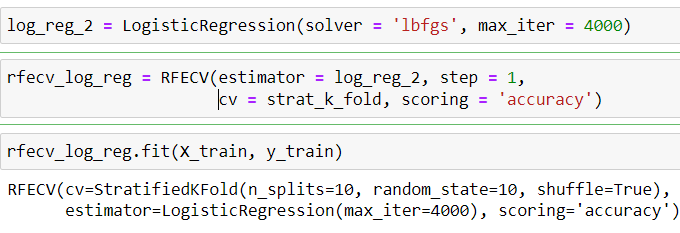
* **Initial Model:**

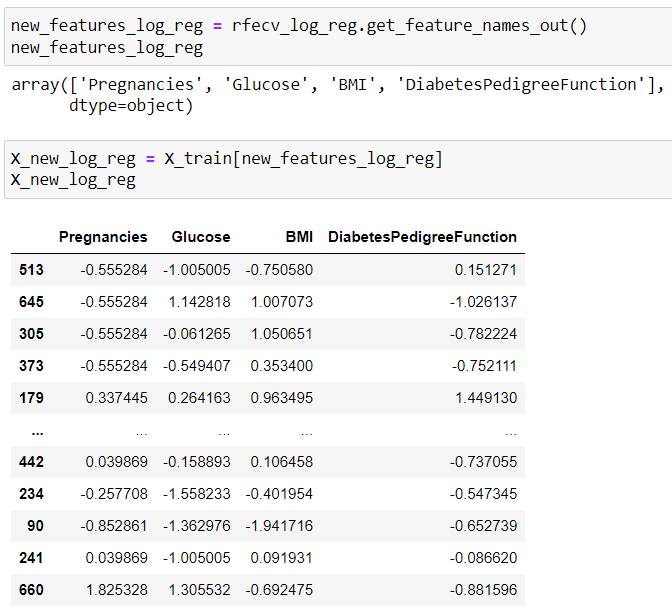


* **Initial Score:**



* **Dimensionality Reduction:**

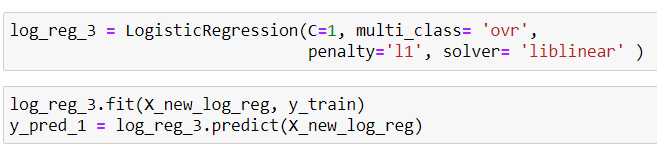




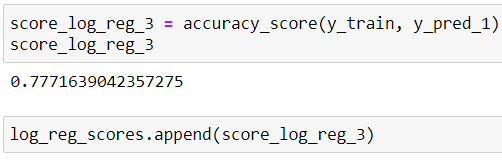
* **Hyper Parameter Tuning:**



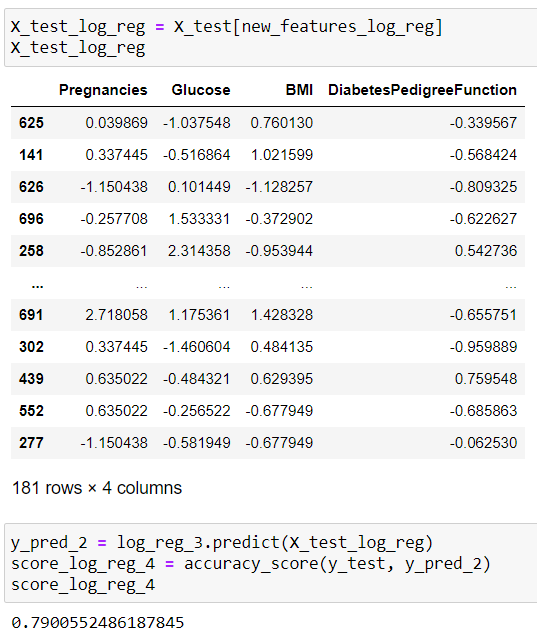
* **Final Model:**



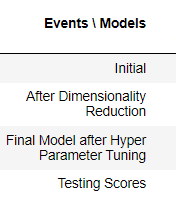
* **Training Scores:**



* **Testing:**



* **Final Scores:**

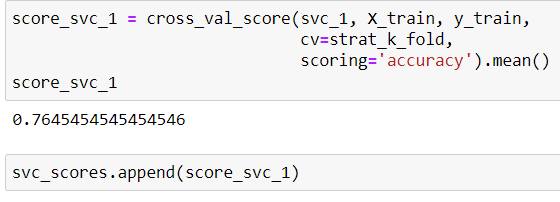
****

1. **Support Vector Classifier-**

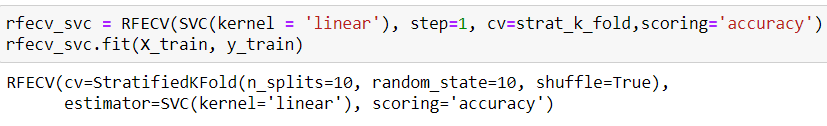
* **Initial Model:**

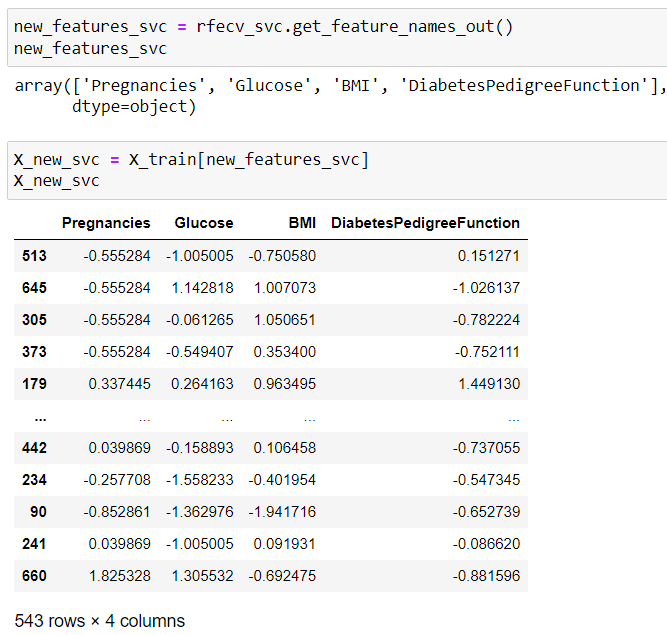


* **Initial Score:**

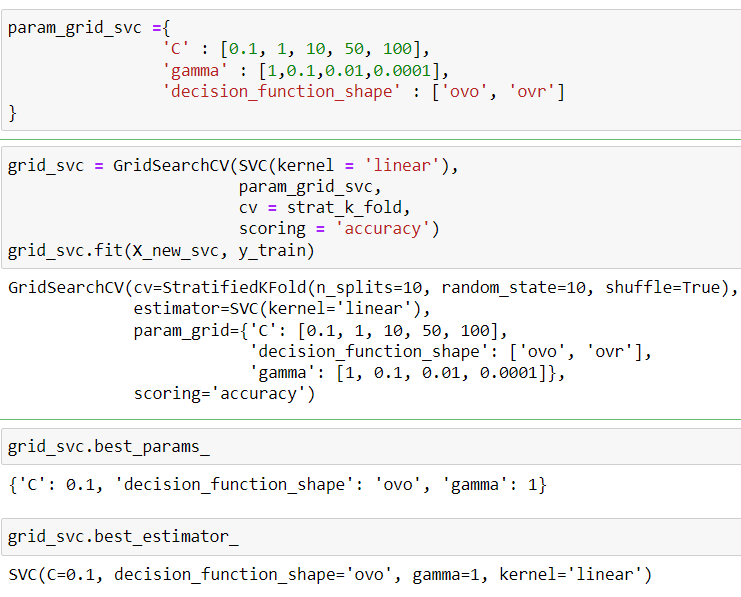


* **Dimensionality Reduction:**

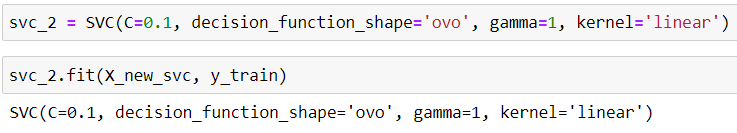




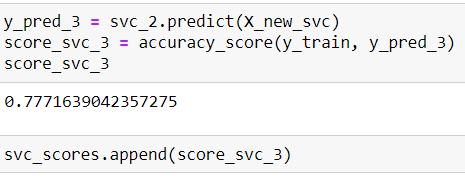
* **Hyper Parameter Tuning:**



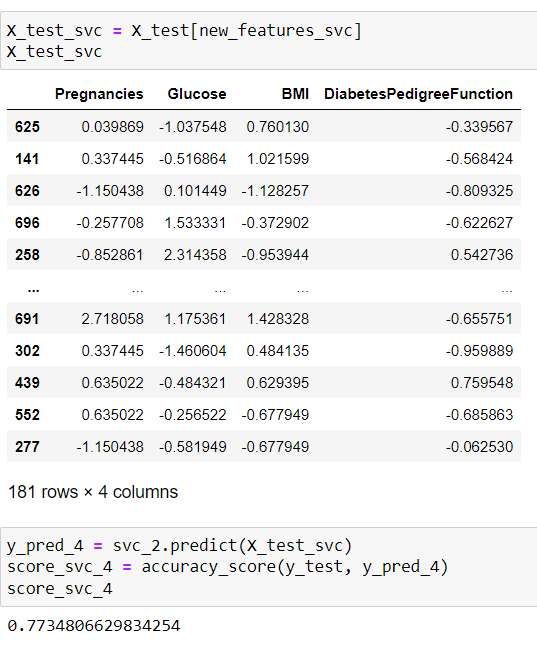
* **Final Model:**



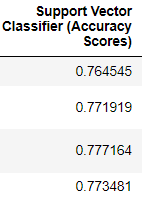
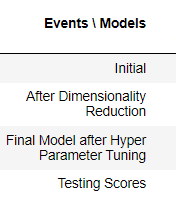
* **Training Scores:**



* **Testing:**



* **Final Scores:**

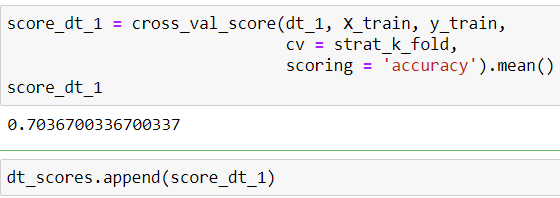
****

1. **Decision Tree Classifier-**

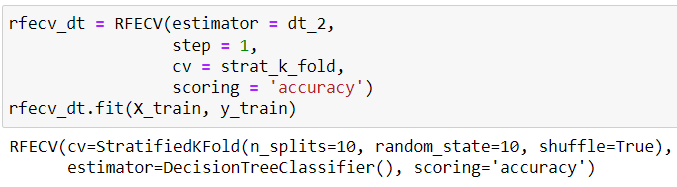
* **Initial Model:**

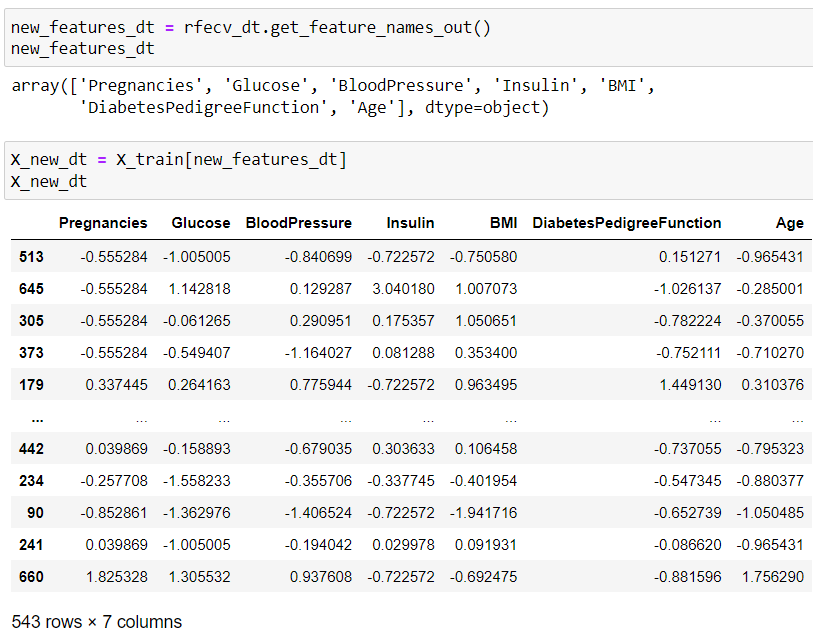


* **Initial Score:**



* **Dimensionality Reduction:**

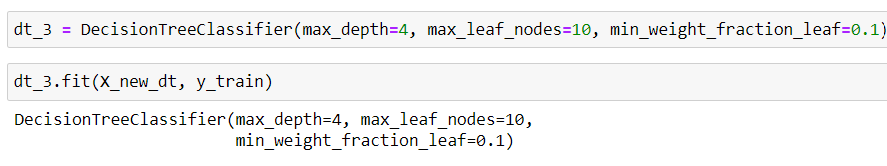




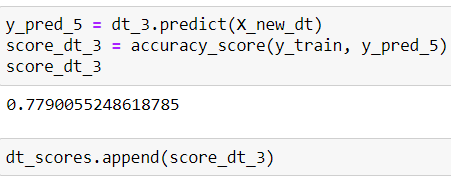
* **Hyper Parameter Tuning:**



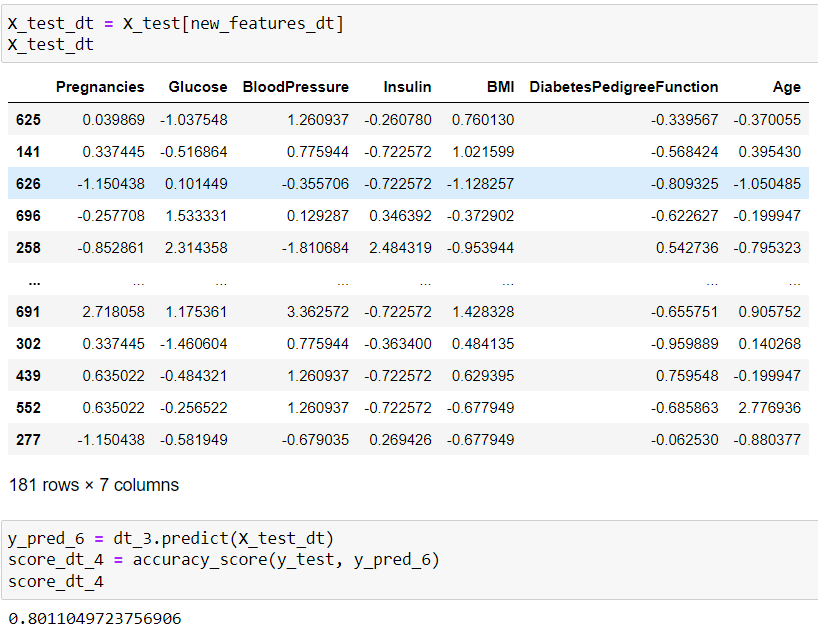
* **Final Model:**



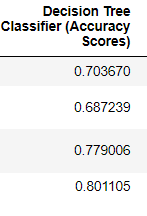
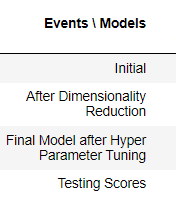
* **Training Scores:**



* **Testing:**



* **Final Scores:**

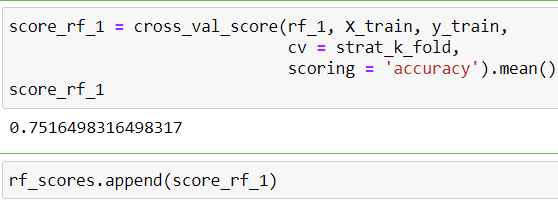
****

1. **Random Forest Classifier-**

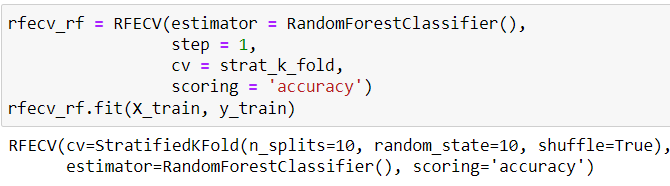
* **Initial Model:**

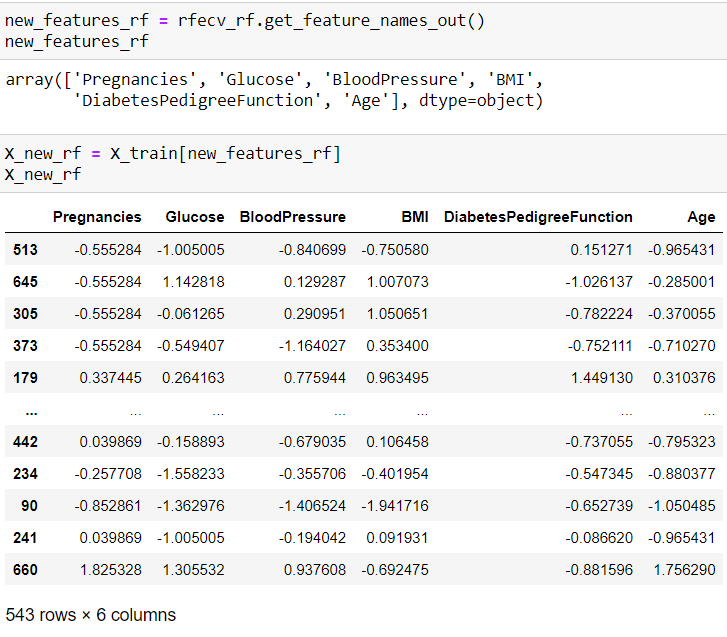


* **Initial Score:**

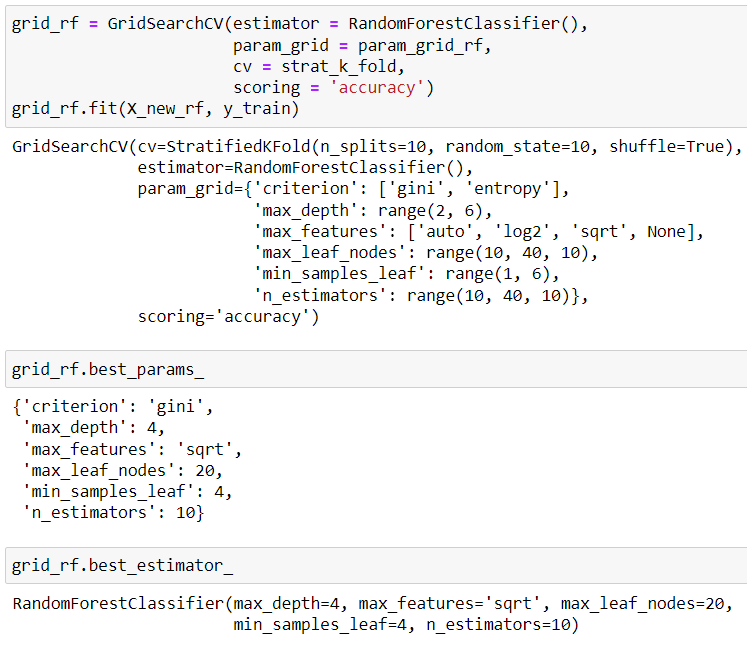


* **Dimensionality Reduction:**

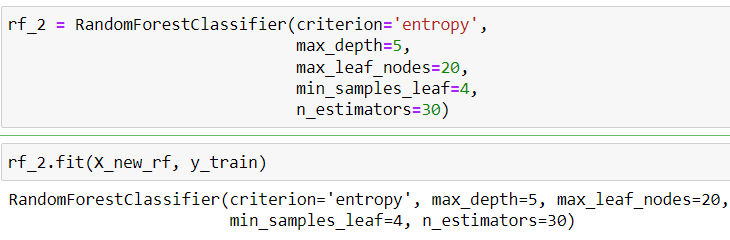




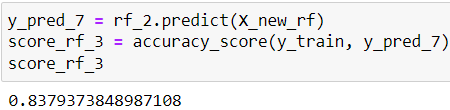
* **Hyper Parameter Tuning:**



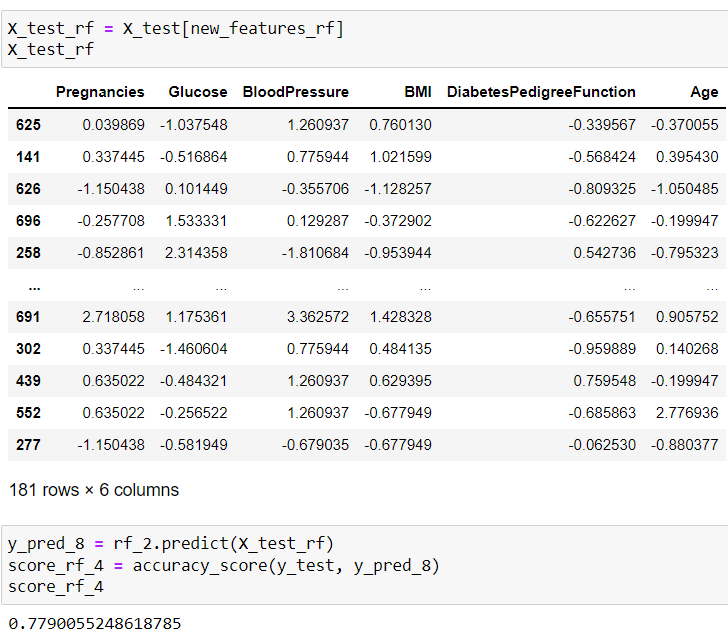
* **Final Model:**



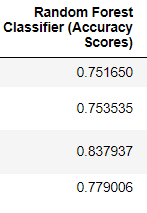
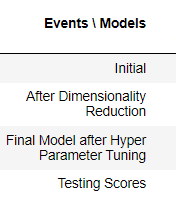
* **Training Scores:**



* **Testing:**



* **Final Scores:**

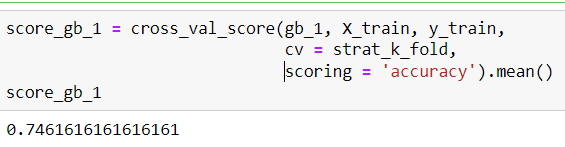
****

1. **Gradient Boosting Classifier-**

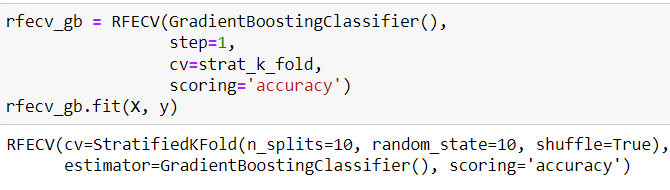
* **Initial Model:**

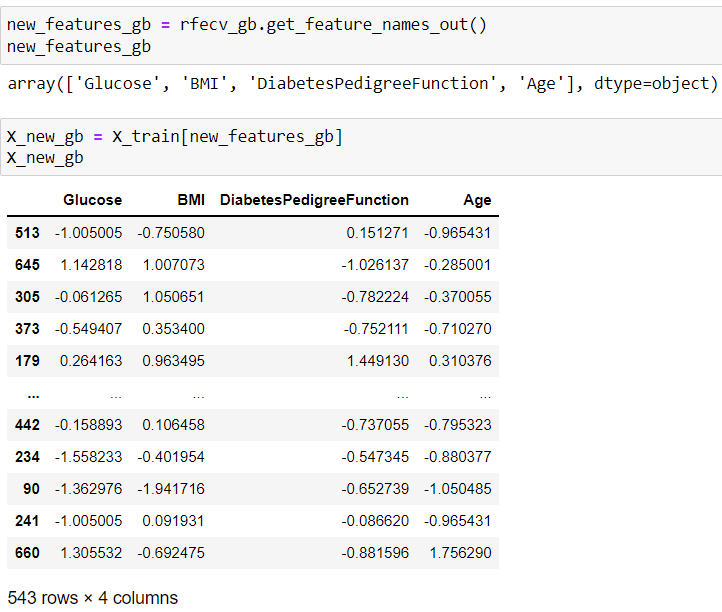


* **Initial Score:**

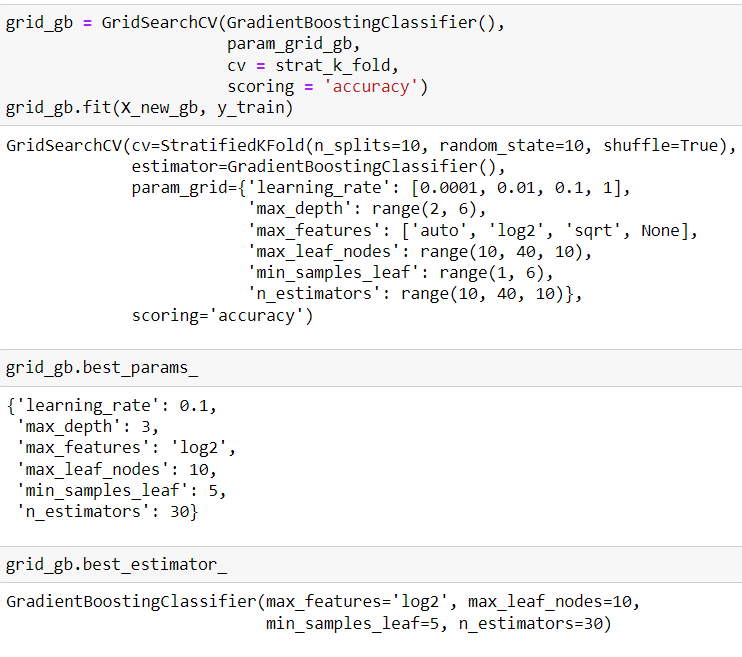


* **Dimensionality Reduction:**

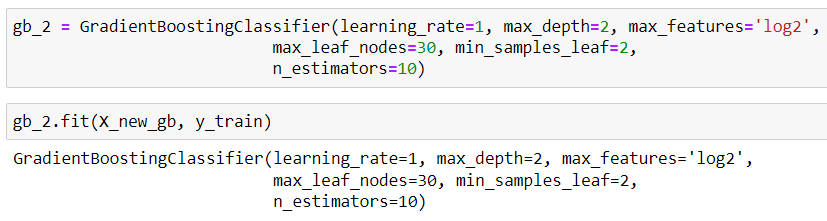




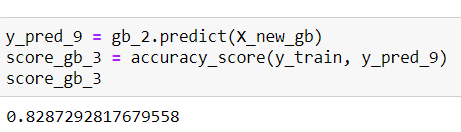
* **Hyper Parameter Tuning:**



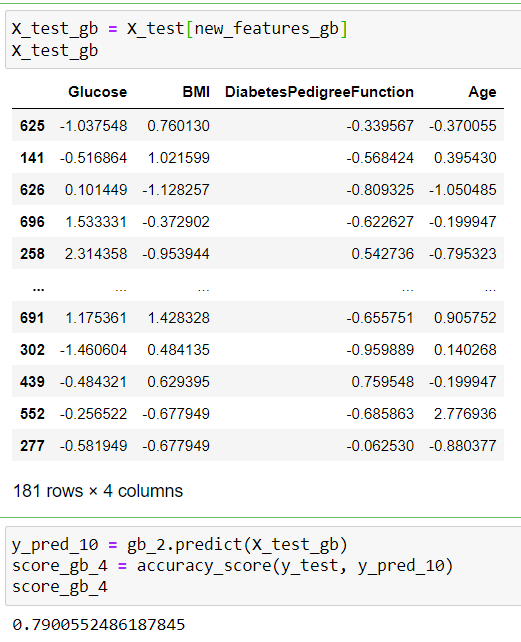
* **Final Model:**



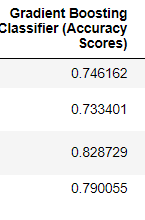
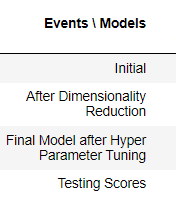
* **Training Scores:**



* **Testing:**



* **Final Scores:**

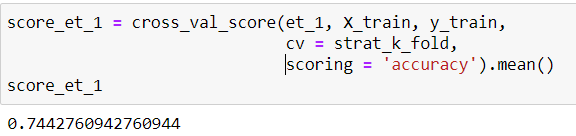
****

1. **Extra Trees Classifier-**

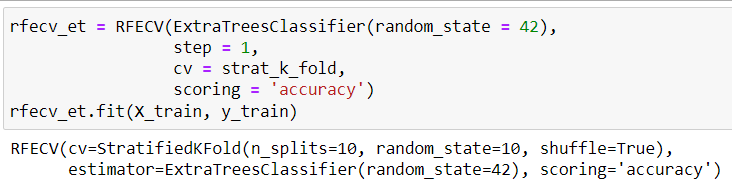
* **Initial Model:**

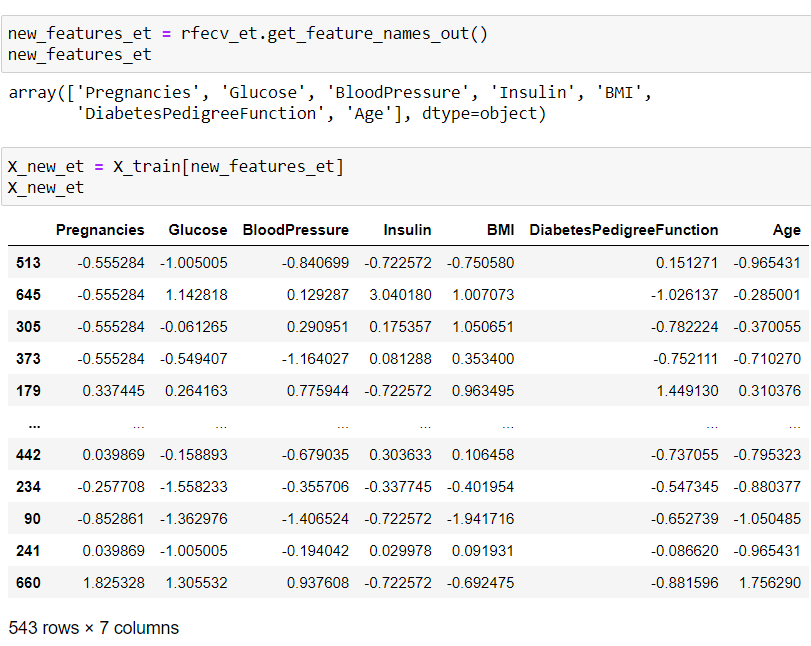


* **Initial Score:**

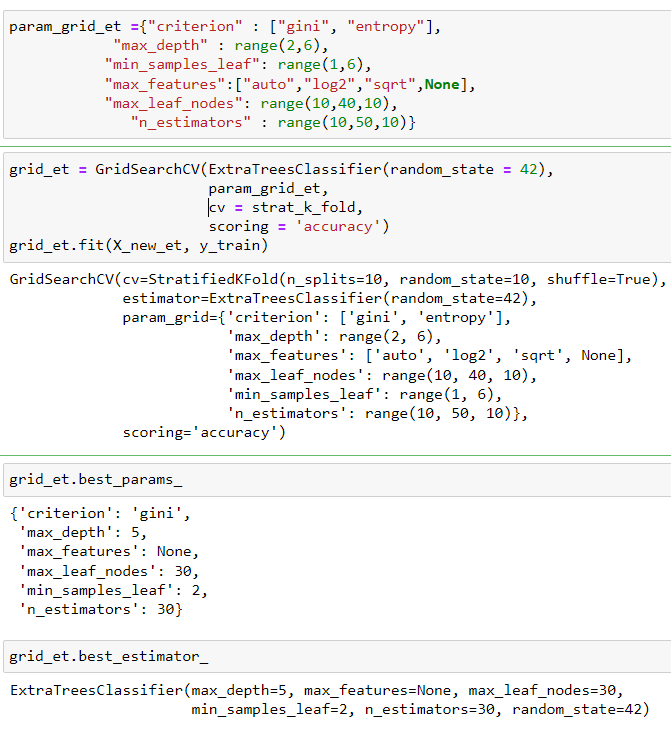


* **Dimensionality Reduction:**

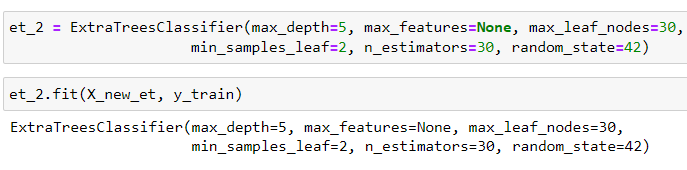




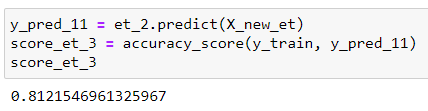
* **Hyper Parameter Tuning:**



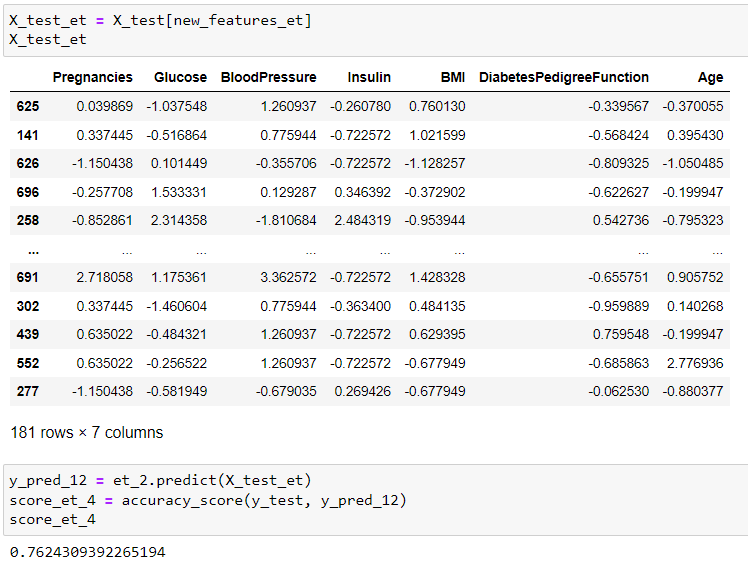
* **Final Model:**



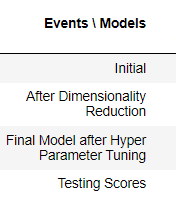
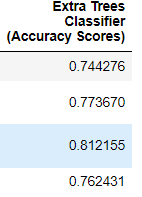
* **Training Scores:**



* **Testing:**

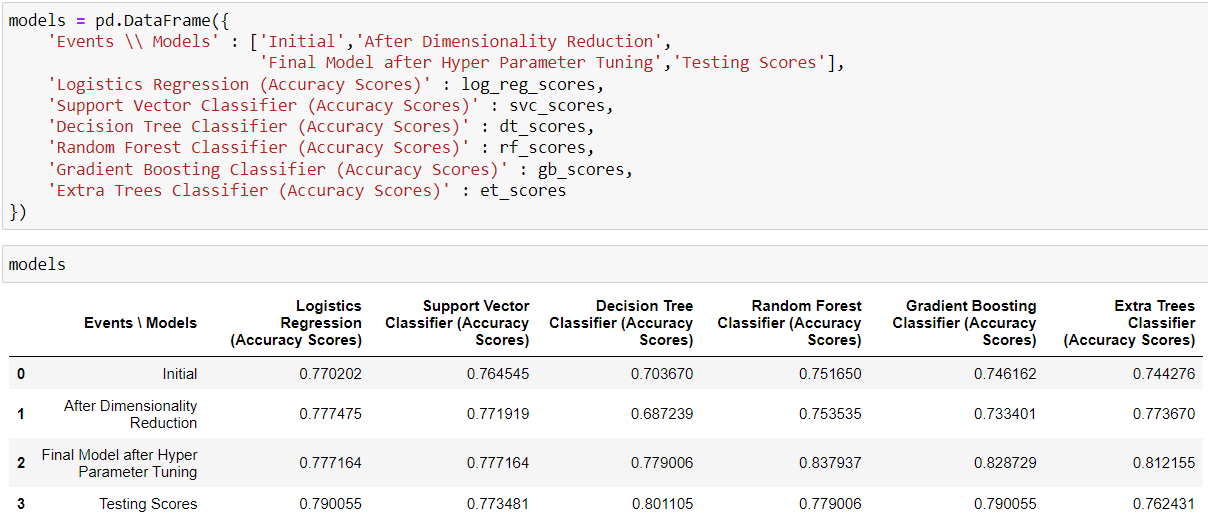


* **Final Scores:**

** **

**Conclusion**

Thus, by enlisting each model considered and the corresponding Accuracy Scores.



Decision Tree Classifier performed very well on training data with accuracy of 0.779006.

And the performed extremely well on testing with accuracy of 0.801105.

Logistic Regression seems as good as logistic regression with training score of 0.777164 and testing score of 0.790055

Thus Decision Tree Classifier is the best model to be chosen for prediction of presence of Diabetes based on data by **National Institute of Diabetes and Digestive and Kidney Diseases.**

**-Thank You-**