# 

# Report on Early Stage Diabetes Risk Prediction

# Exposys Data Labs

Domain of the project: Data Science

**Data Science**

**Given Problem Description:**

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 accordingly to either prevent the occurrence of this disease or control the disease (For people who already have the disease)

**Task:**

1. Prepare the data-set using several methods to train the model
2. Build model which can give high accuracy of predicting the disease

**Index:**

|  |  |
| --- | --- |
| **Sno** | **Contents** |
|  | Abstract |
|  | Introduction |
|  | Existing Method |
|  | Proposed method of Architecture |
|  | Implementation |
|  | Conclusion |

**Abstract:**

A chronic disease that is more common in all age groups and many people suffer from is Diabetes. Prediction of this chronic disease in advance or at an early stage can help people take necessary precautions in order to prevent the after effects of the disease.

We are predicting the risks of diabetes whether it is present or not, early by considering the following features which has high correlation with the occurrence of diabetes. Features are Age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity.

The algorithms which we used are supervised learning where we loaded both input and output data followed by training and testing the data according to respective algorithms. The algorithms which we used are Naïve Bayes, Random forest, Decision tree, Support vector machine, Logistic regression, KNN. The comparison is made among all these ML algorithms and we chose the models which gave highest accuracy and predicted our output.

**Introduction:**

Diabetes is a disease whereby blood sugar (glucose) is not metabolized in the body. This increases the glucose in the blood to alarmingly high levels. This is known by the name hyperglycemia. In this condition, body is unable to produce sufficient insulin. The other possibility is that body cannot respond to the produced insulin. Diabetes is incurable; it has to be controlled. A diabetic person can develop severe complications like nerve damage, heart attack, kidney failure and stroke. According to statistics in 2017, an estimated 8.8% of global population has diabetes. This is likely to increase to 9.9% by year 2045.

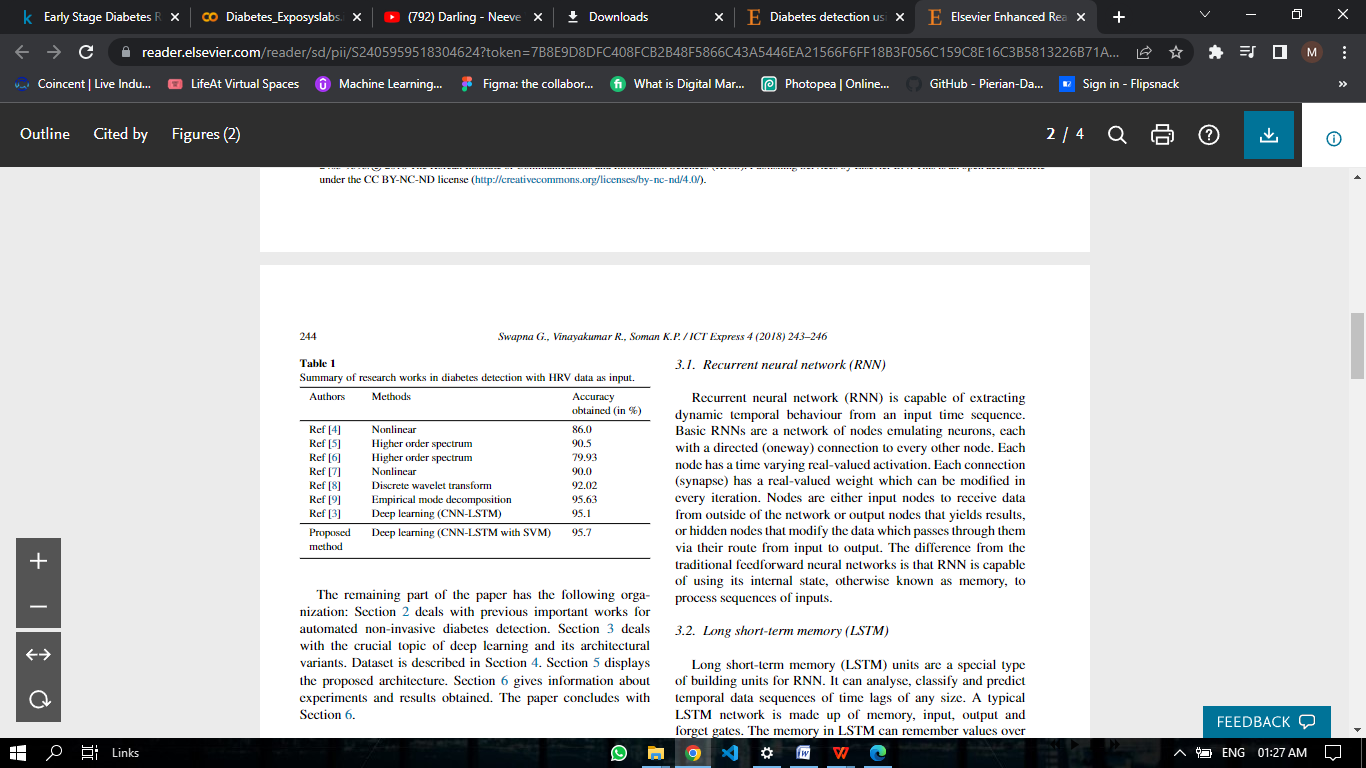
Diabetes is one of the fastest growing chronic life threatening diseases that have already affected 422 million people worldwide according to the report of World Health Organization (WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase. The early diagnosis of diabetes is only possible by proper assessment of both common and less common sign symptoms, which could be found in different phases from disease initiation up to diagnosis. To predict the likelihood of having diabetes requires a dataset, which contains the data of newly diabetic or would be diabetic patient. In this work, we have used such a dataset of 520 instances, which has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh.

**Existing Method:**

****

Accuracy obtained in this previous research work is:

Summary of research works in diabetes detection with HRV data as input

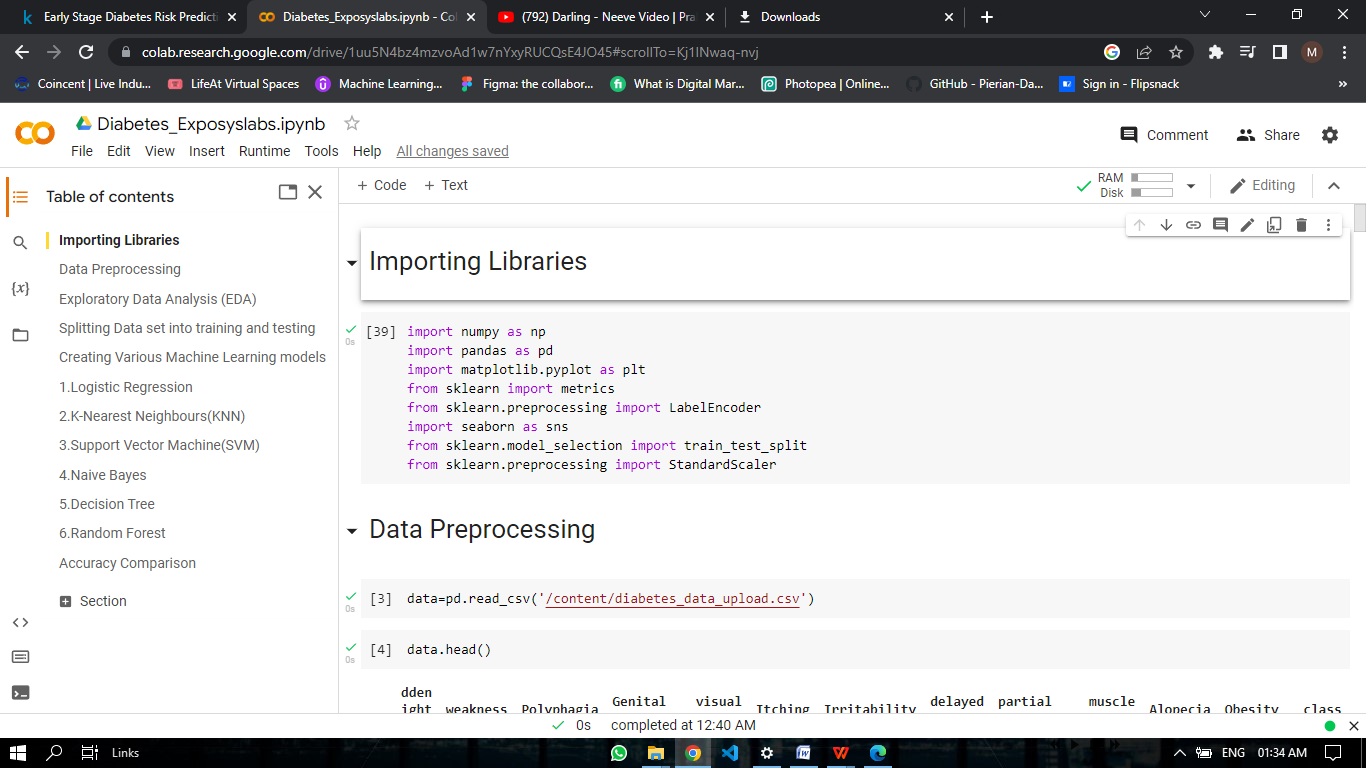
****

Maximum accuracy obtained in this paper using deep learning methods is 95.7%.

**Proposed Method of Architecture:**

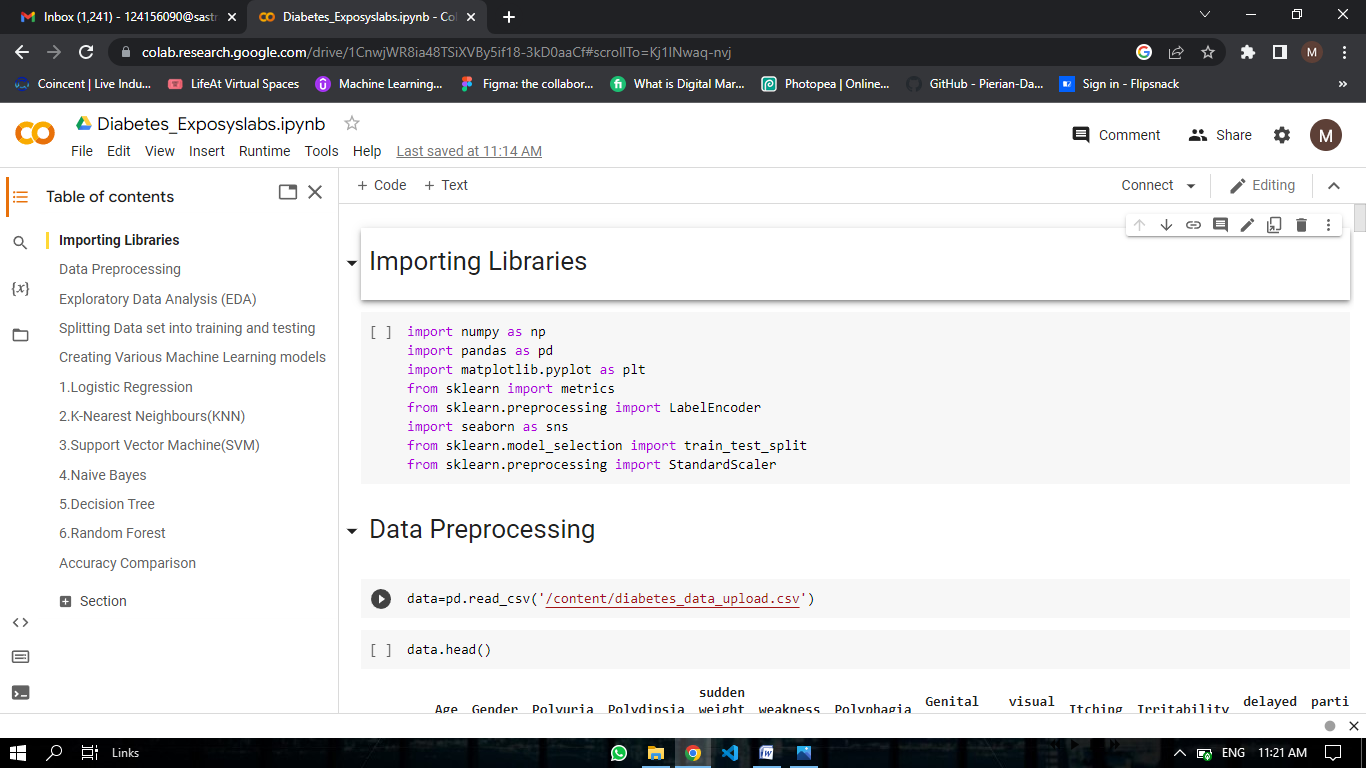
Compare different Machine learning models and then select the best one among them which has the highest accuracy

Workflow:



**Implementation:**

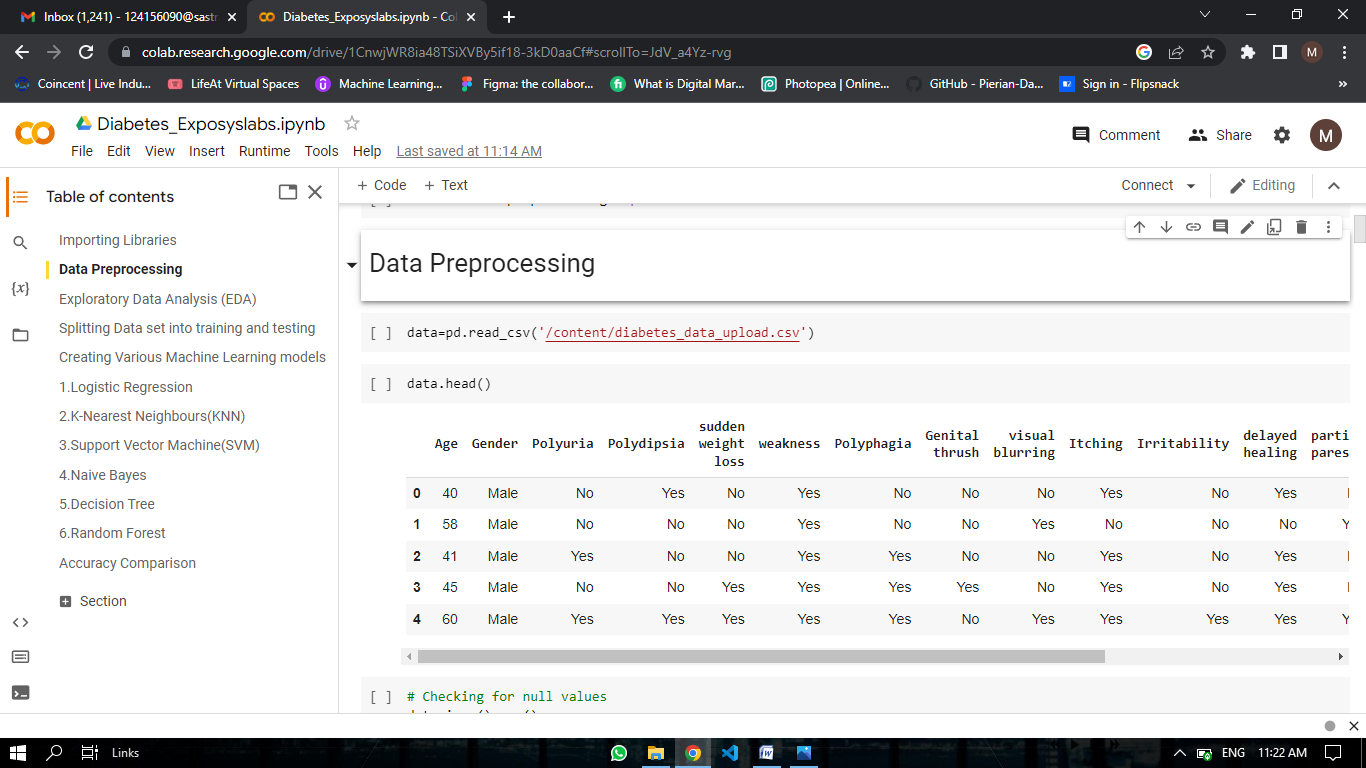
1. Importing libraries



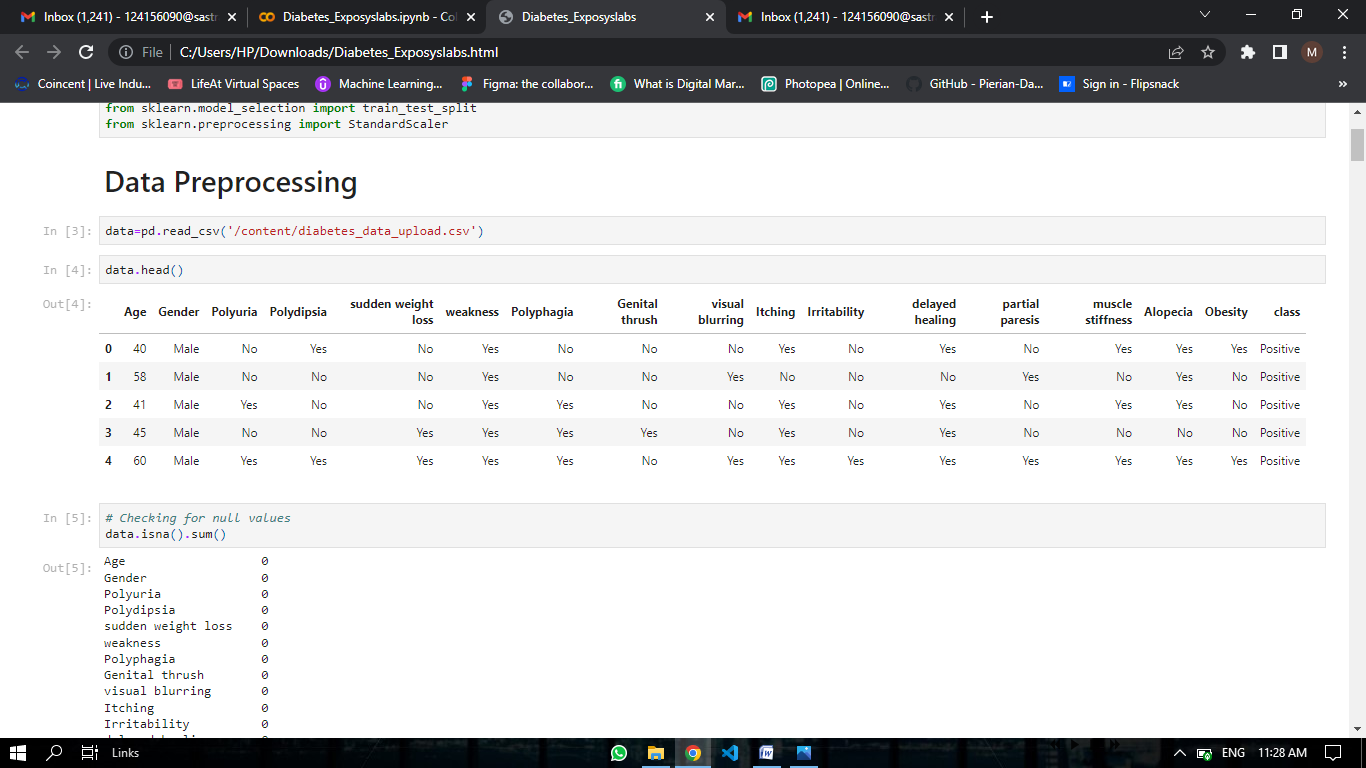
1. Data preprocessing

* Data loading

The dataset is obtained from Kaggle platform

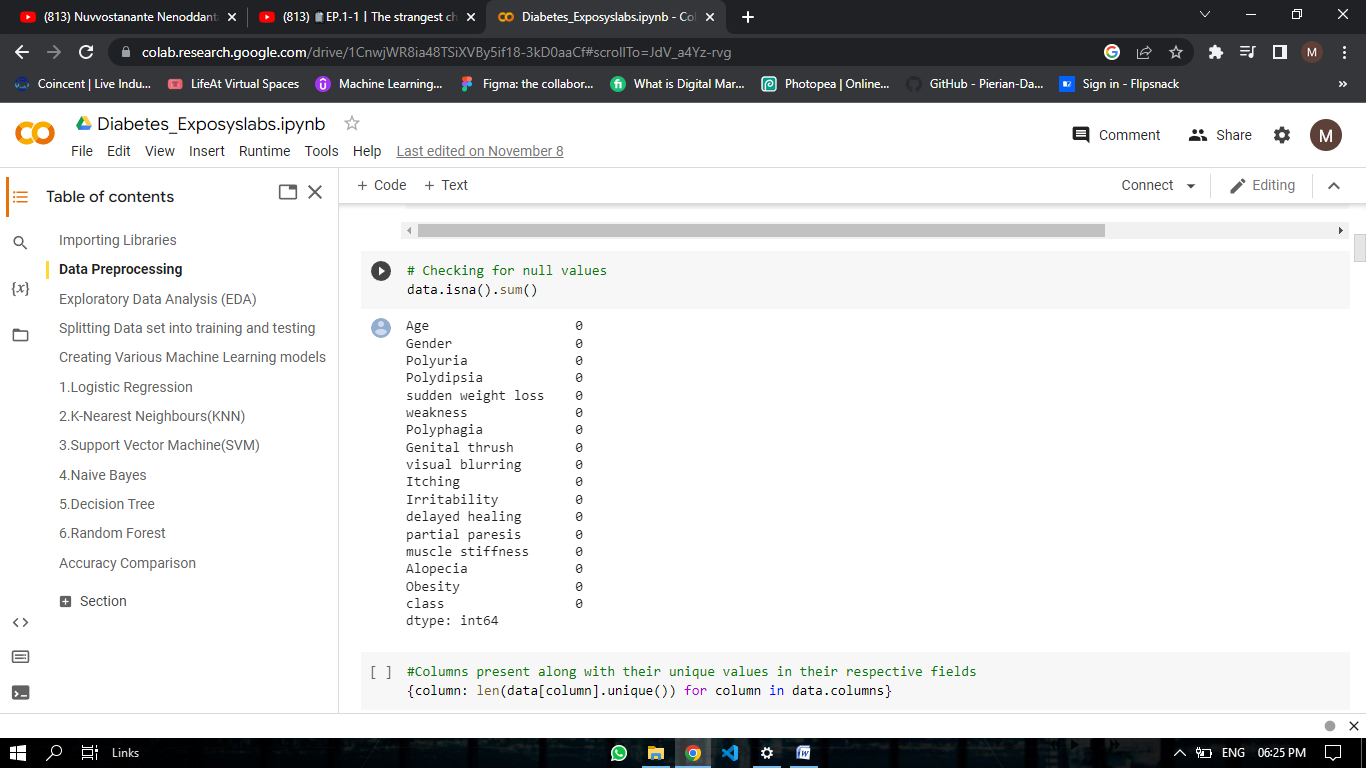
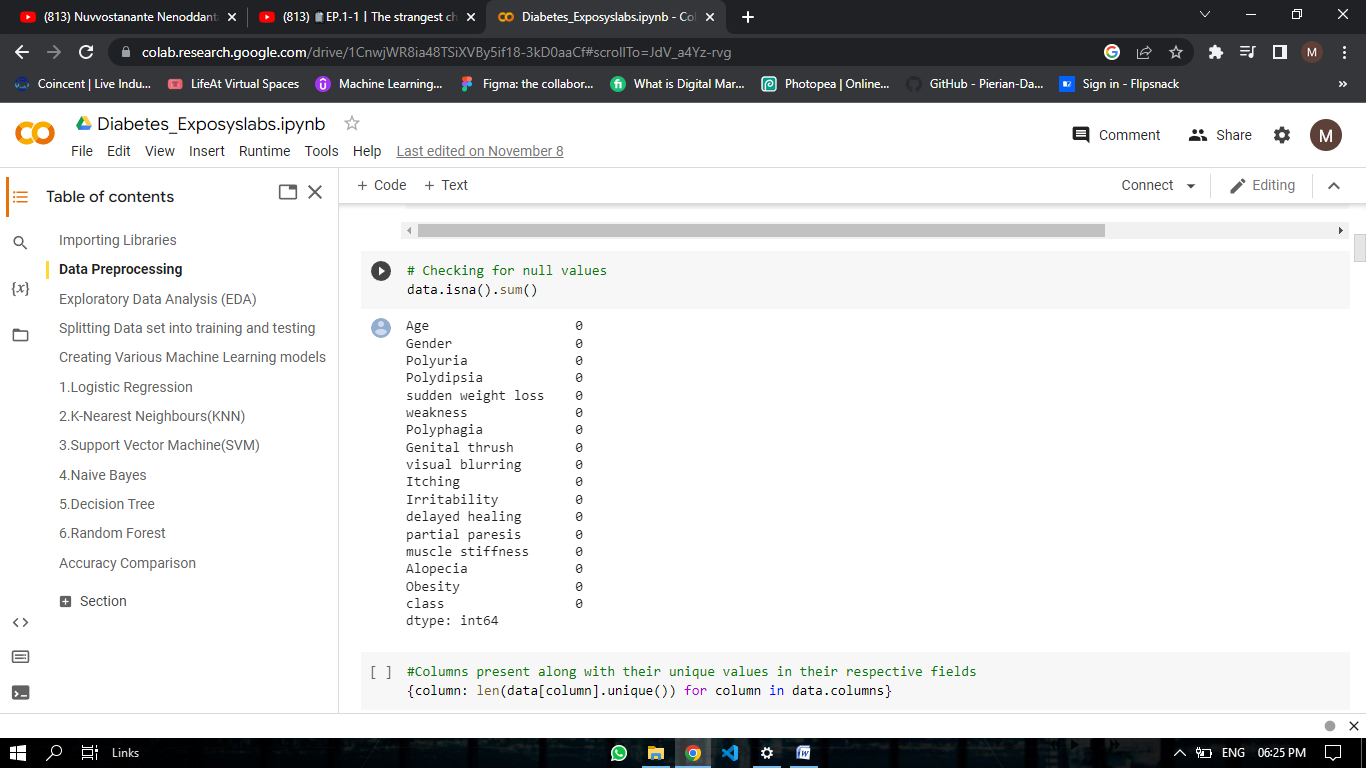


* Learning about data

Inferences drawn from this dataset:

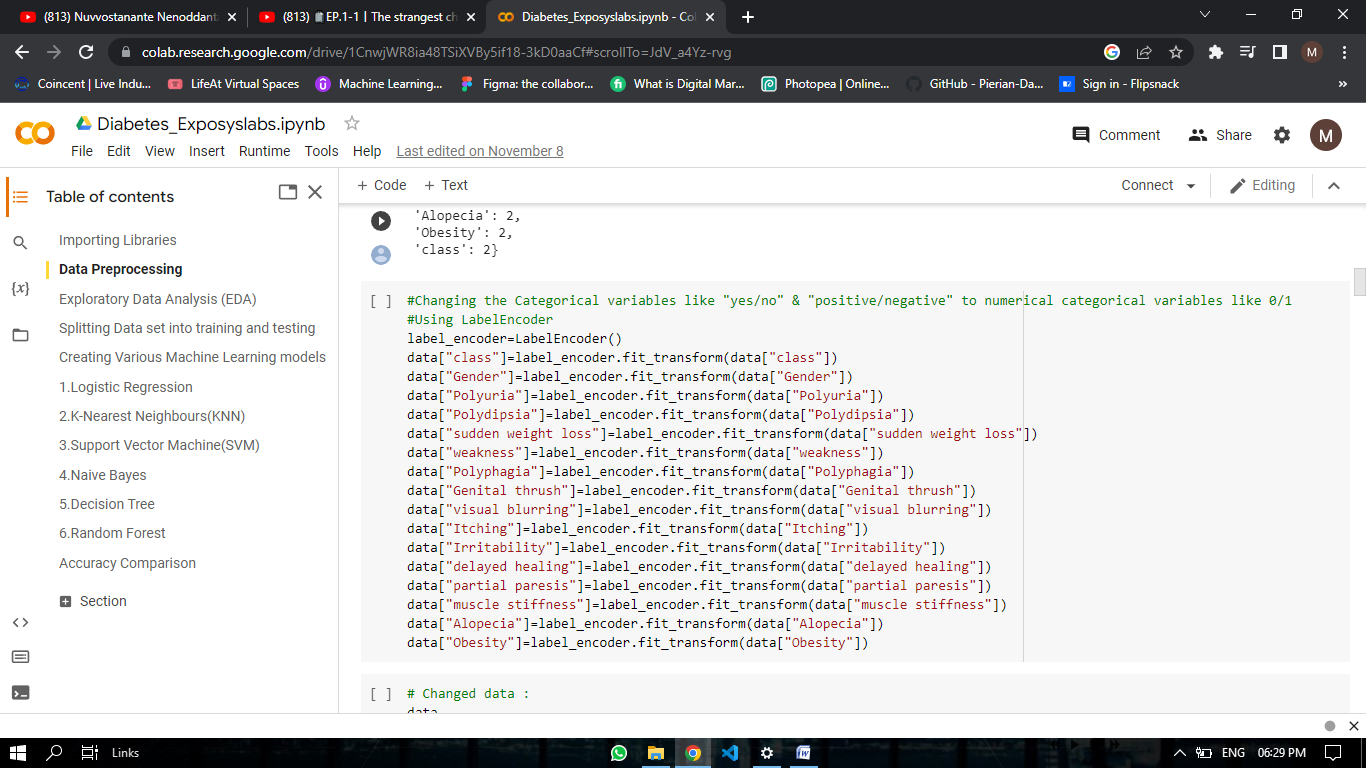
1. Age is a numerical value
2. Columns/Fields except “Age” are all categorical variables “Yes/No” and “Positive/Negative”
3. There are a total of 16 features to predict whether the person has diabetes or not
4. ‘Positive’ indicates that the person has diabetes and ‘negative’ indicates the person doesnot have diabetes

* Checking If the data has any null values



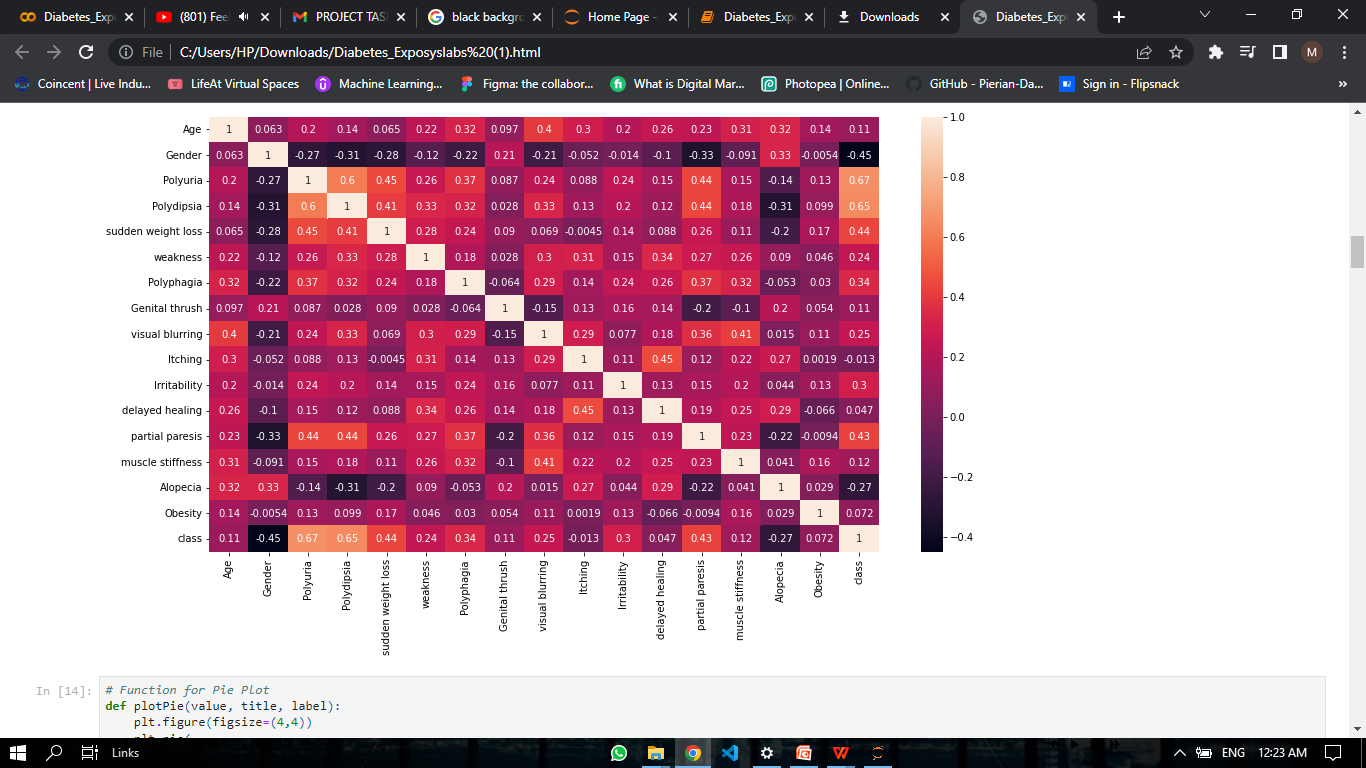
Inferences drawn from this: None of the values are null, so no unknown data is found.

* Converting the categorical variables into 0/1



1. Exploratory data analysis (EDA)

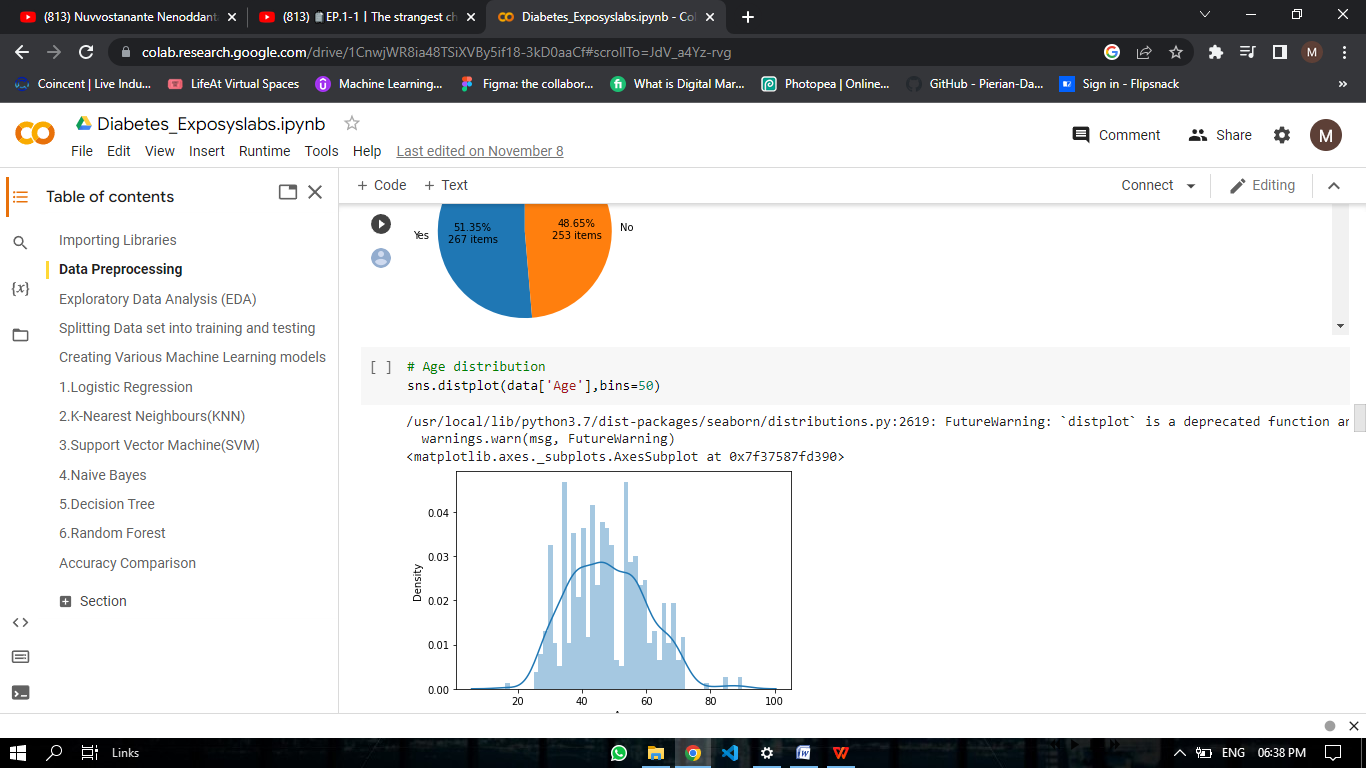
Heat Map:



Inferences drawn from this heat map:

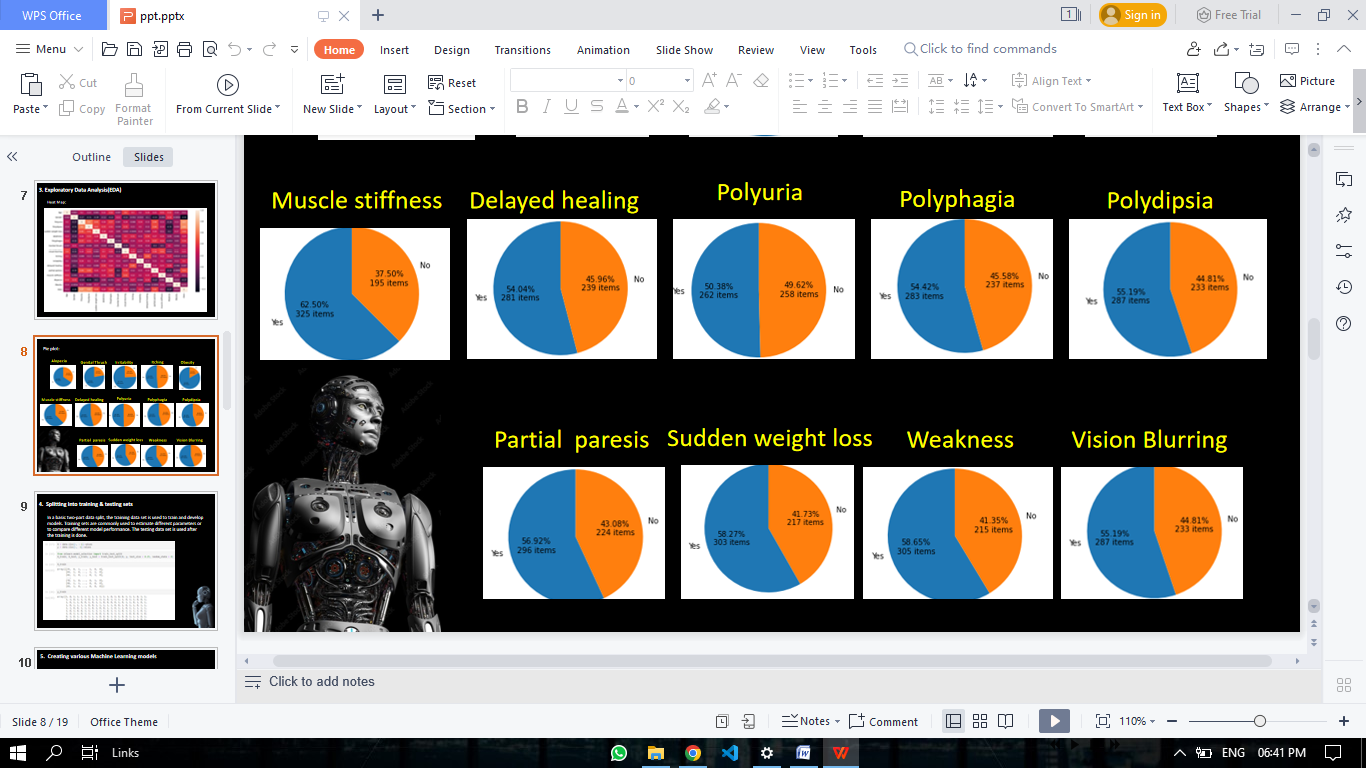
* The corresponding variables for the numbers above 0 are considered positively related to each other
* The variables which have numbers below 0 are considered negatively related to each other
* The last row gives us the relation between the target variable and all the features from the dataset
* Gender (-0.45), Itching (-0.013) and Alopecia(-0.27) are found to be negatively correlated while all the features are positively related with Polyuris (0.67) being the highest and most positively correlated.

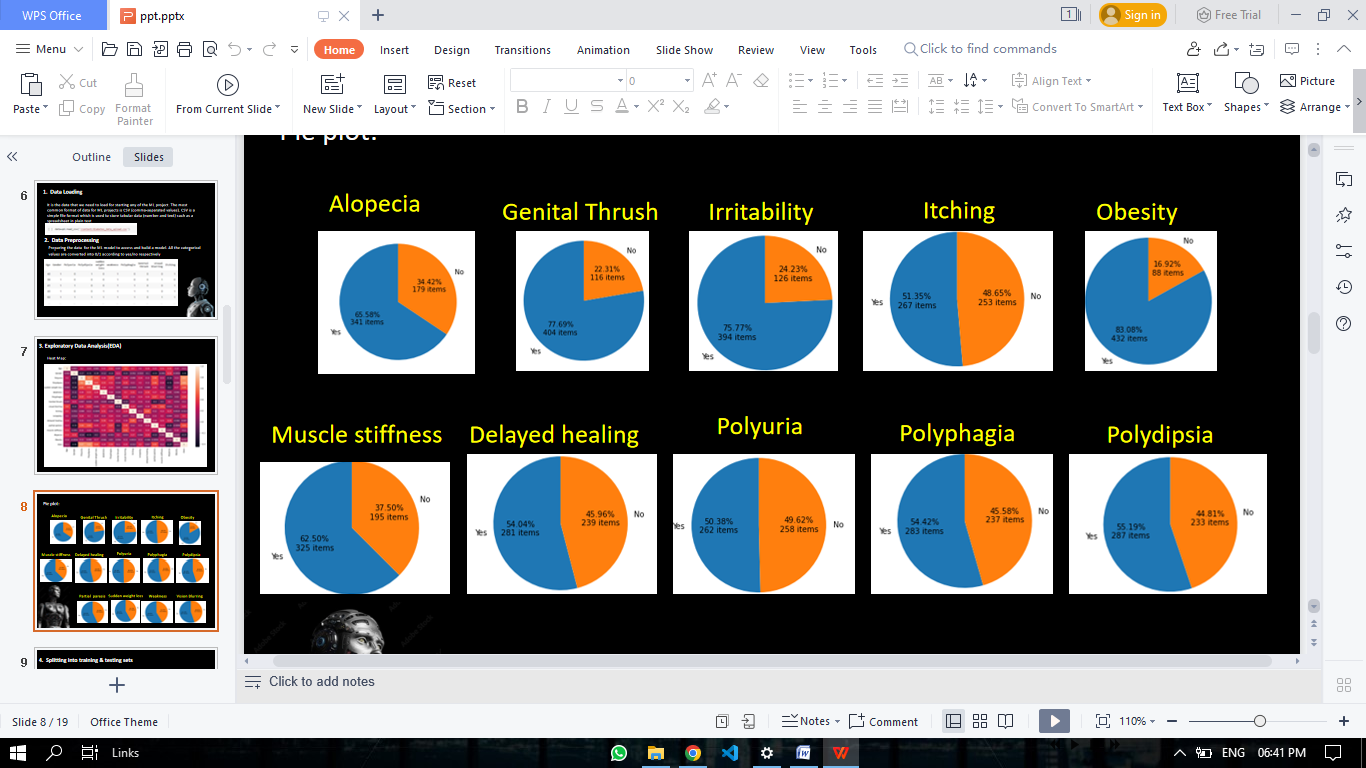
Age Distribution:

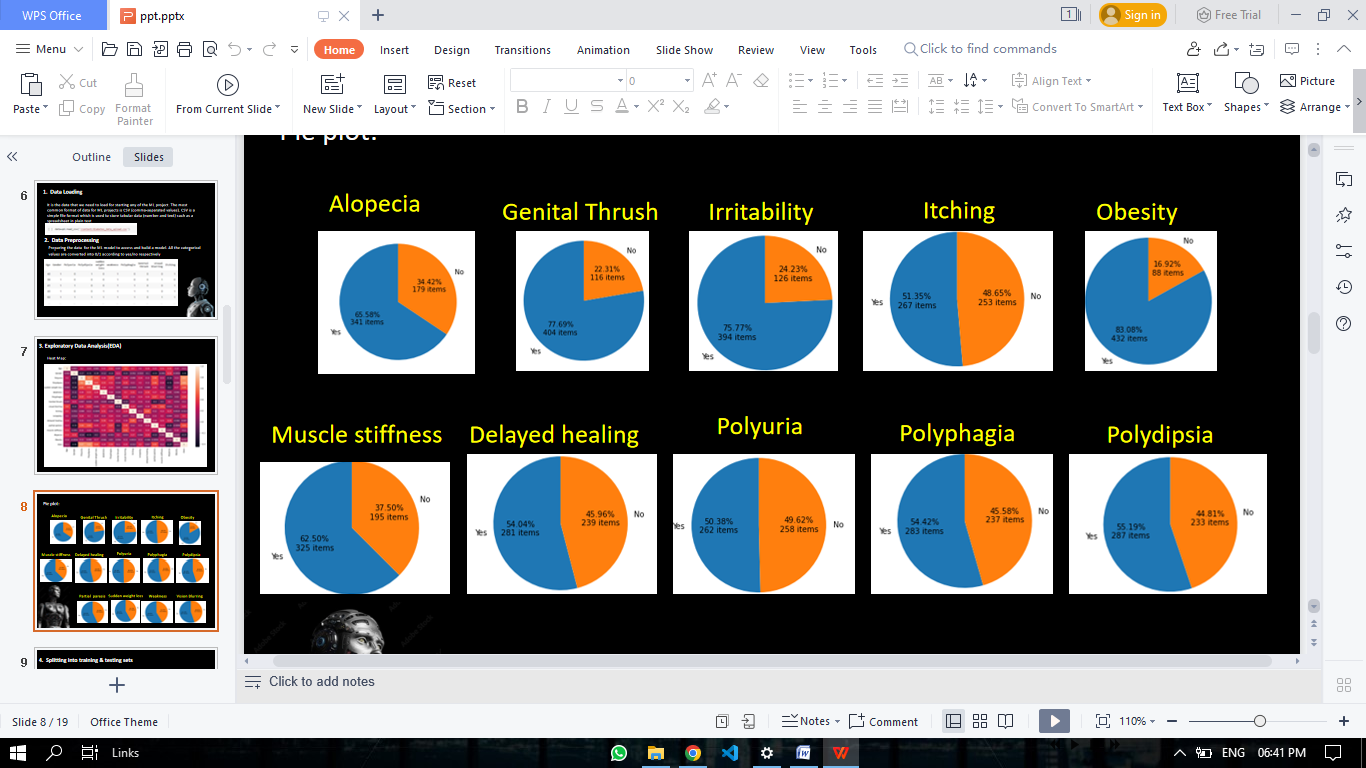


This plot shows us the density against age in this dataser

Pie plot:



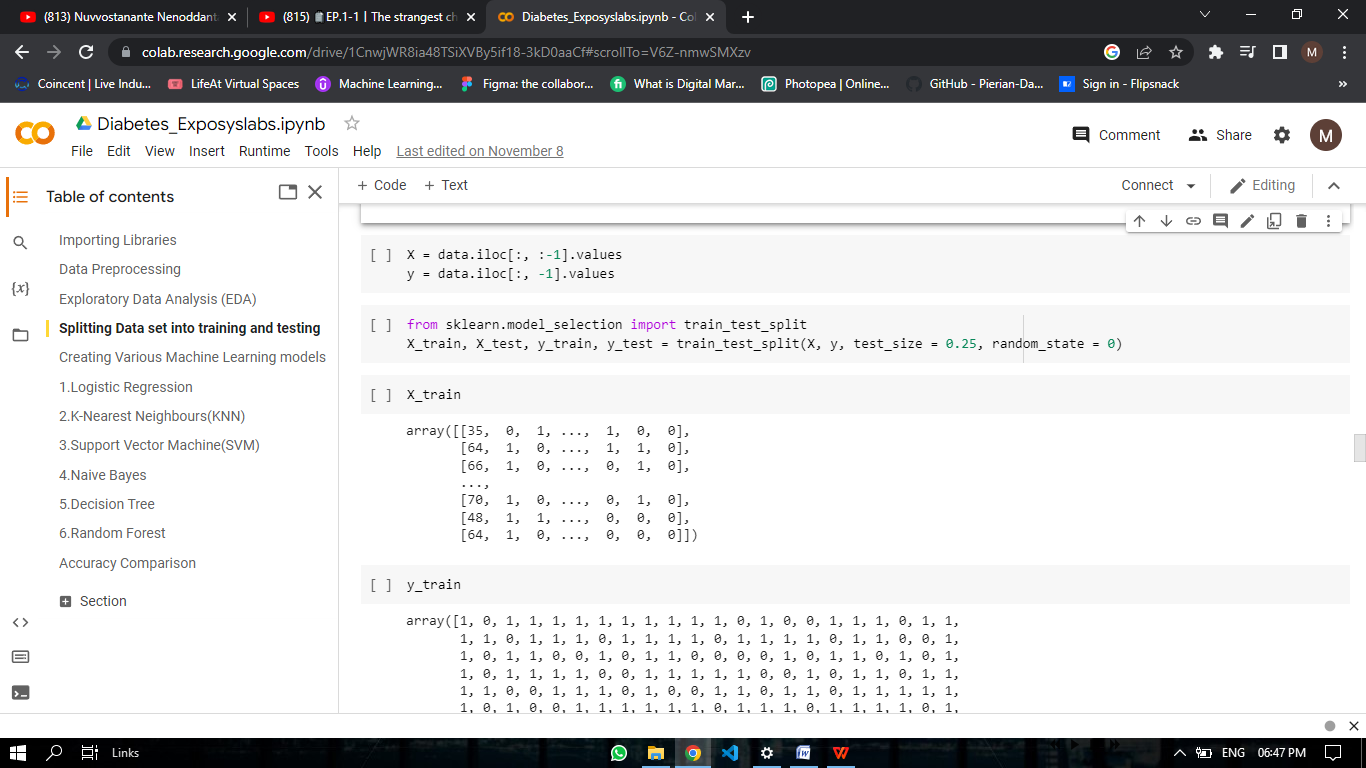




We can find the percentage of yes/no from each feature contributing to the Target variable

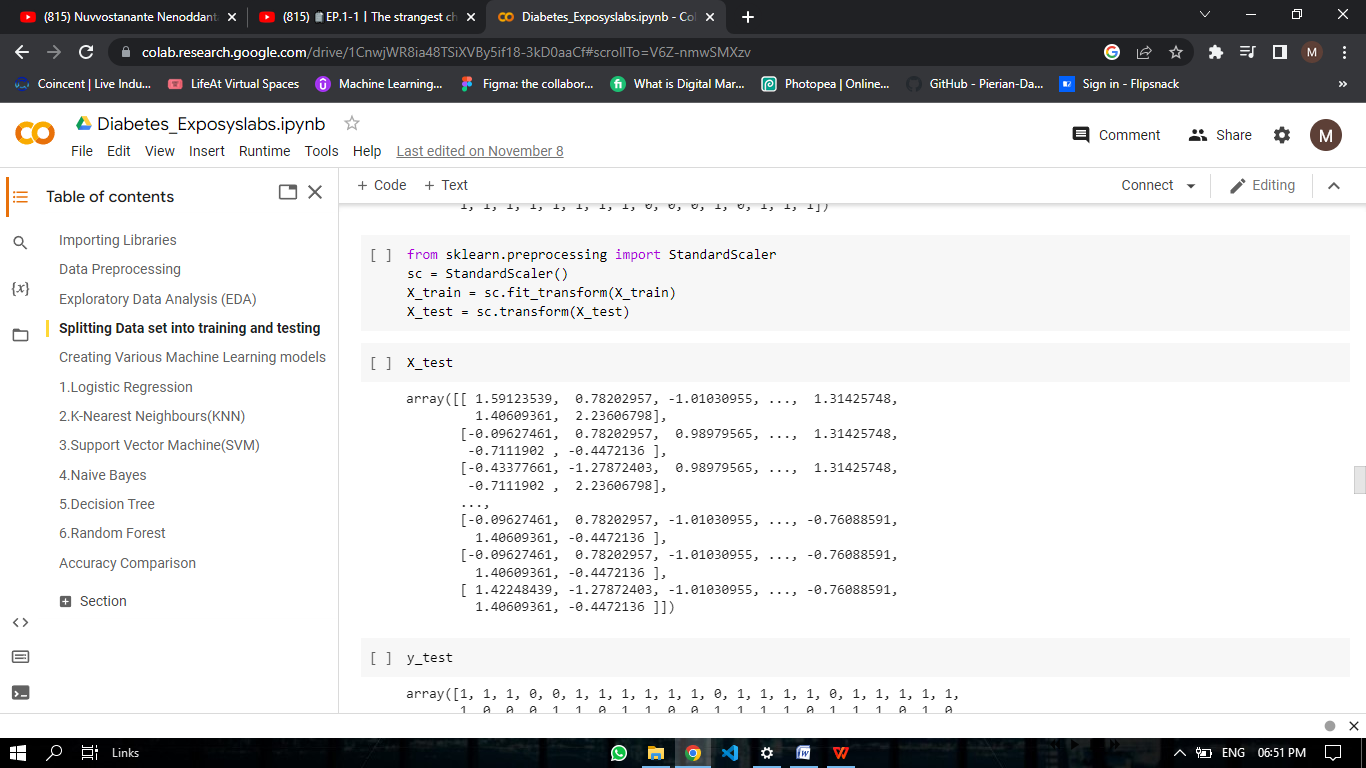
1. Splitting into training and testing data:

Using sklearn we split the data into training and testing set with 25% being the testing set while the rest being the training set



1. Standardizing the Values of the features

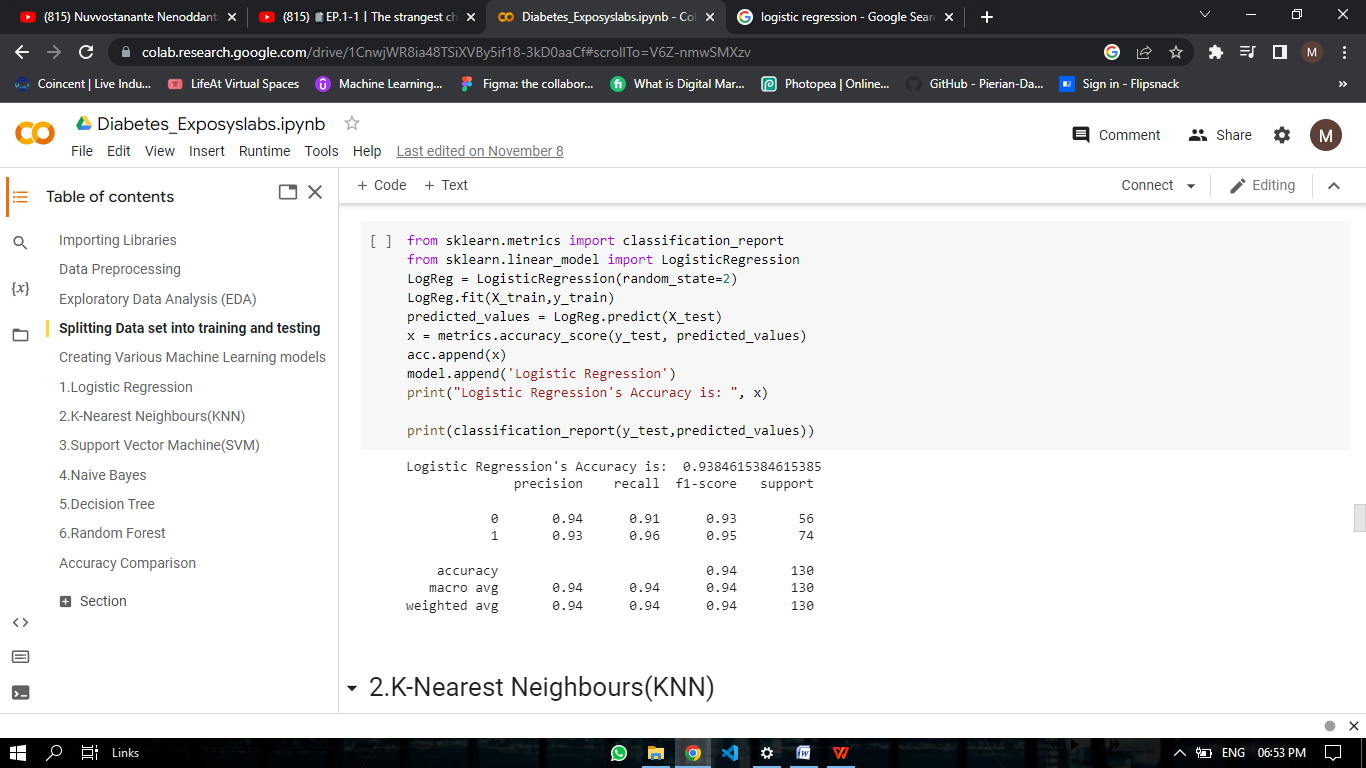
Converting all the data into a standard scaled data of [-2,2], such that the ML models can work o smoother and standard data



1. Creating various ML models

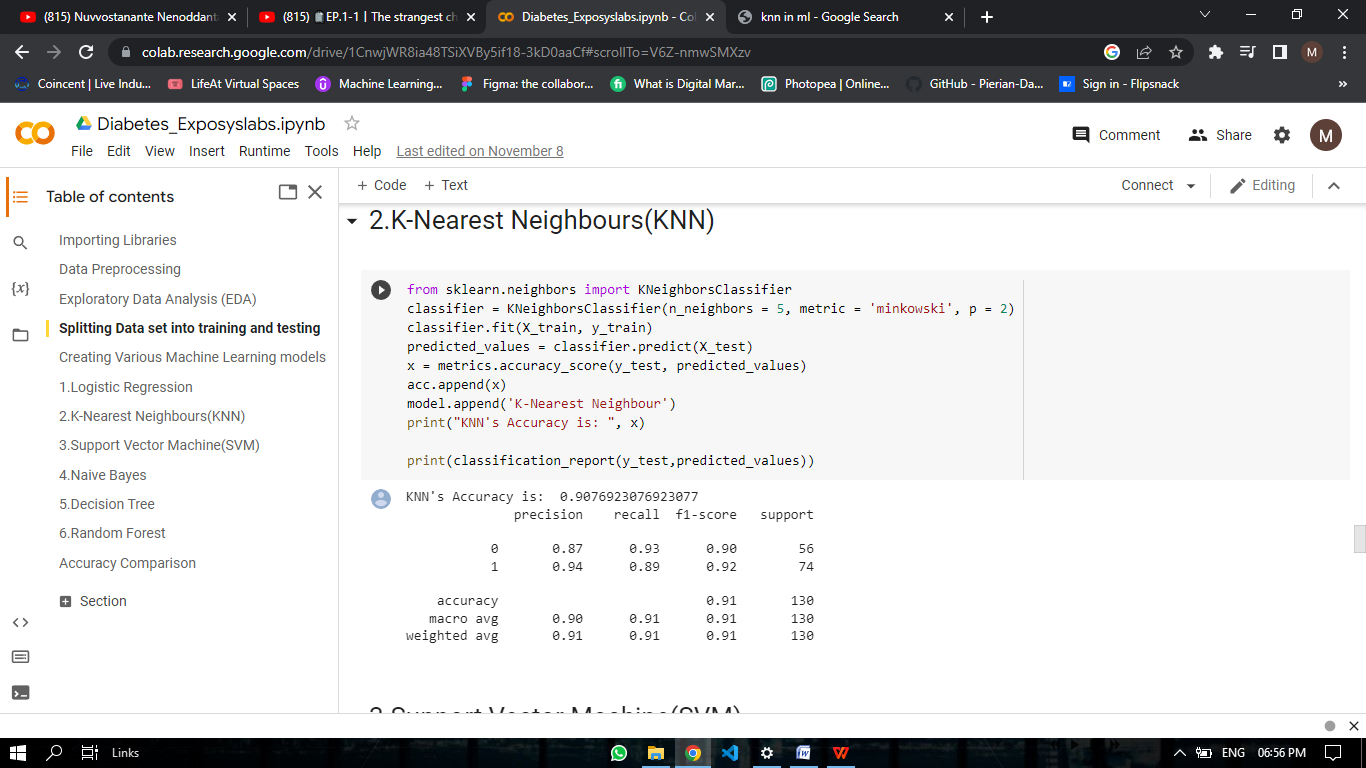
Logistic regression:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.



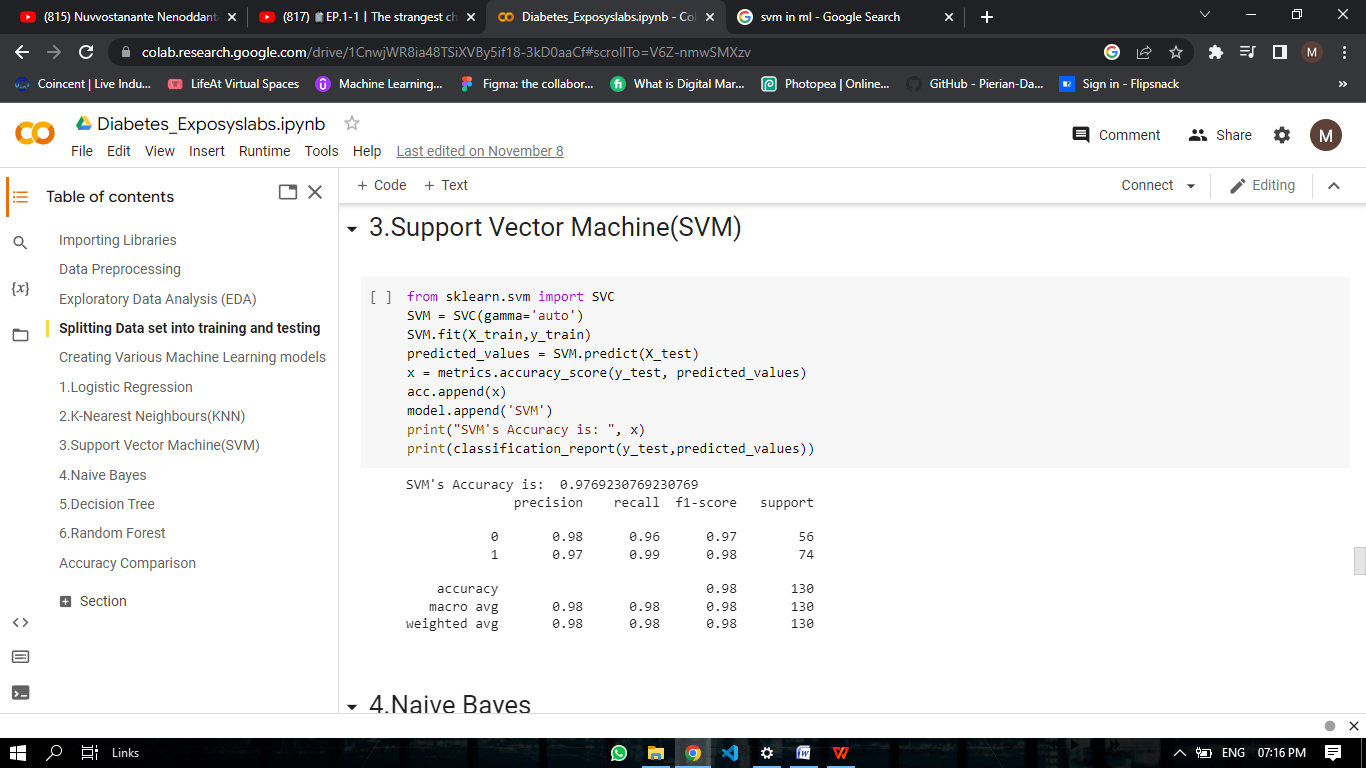
K-Nearest Neighbor:

The k-nearest neighbors classifier (KNN) is a non-parametric supervised machine learning algorithm. It's distance-based: it classifies objects based on their proximate neighbors' classes



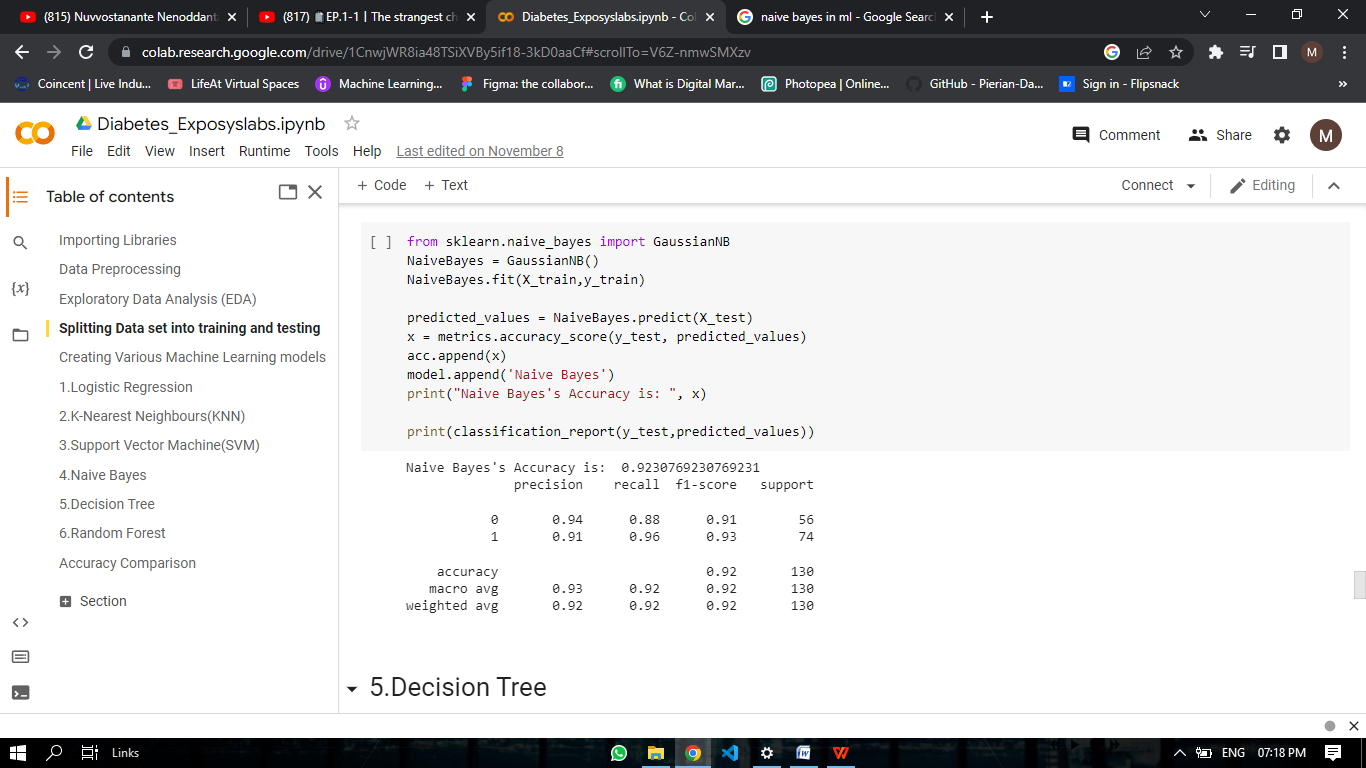
Support Vector Machines (SVM):

It is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyper plane in an N-dimensional space that distinctly classifies the data points.



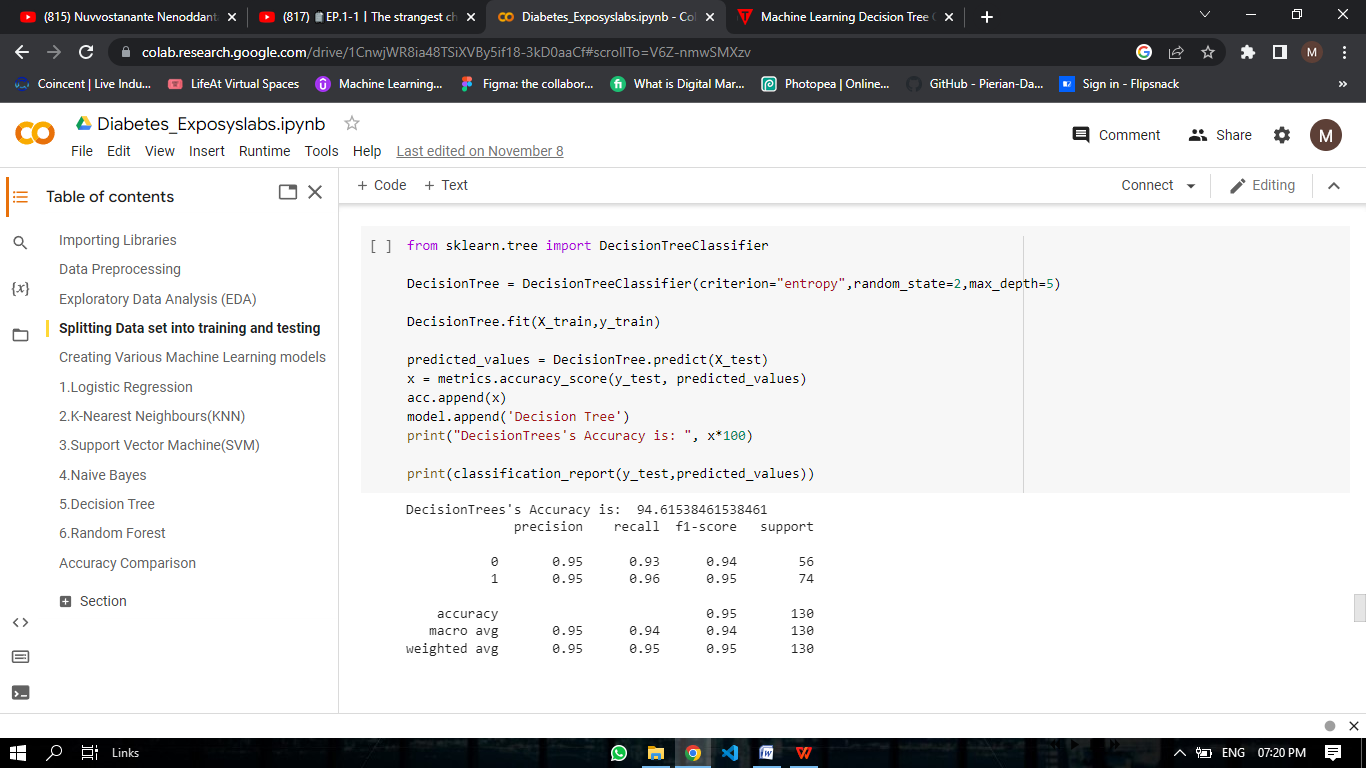
Naïve Bayes:

Naïve Bayes is a simple learning algorithm that utilizes Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy.



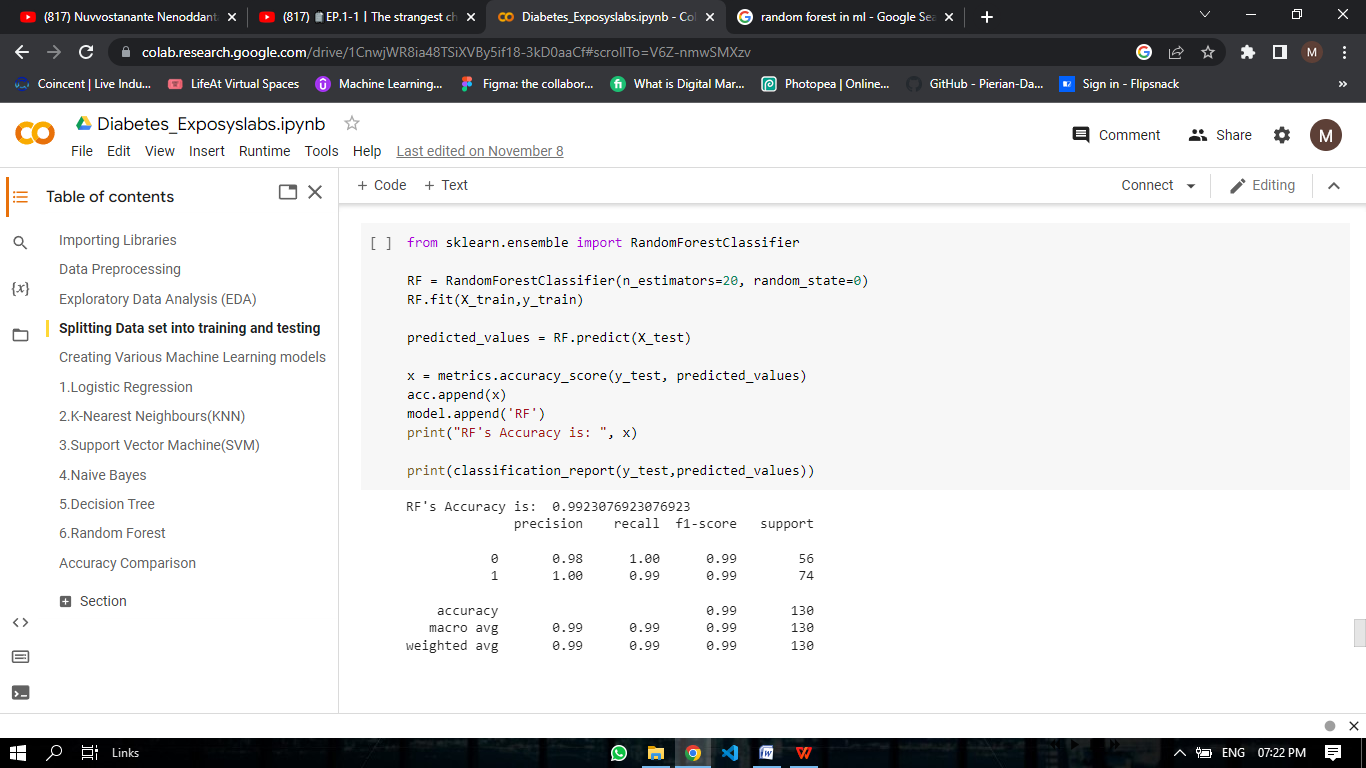
Decision tree:

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.



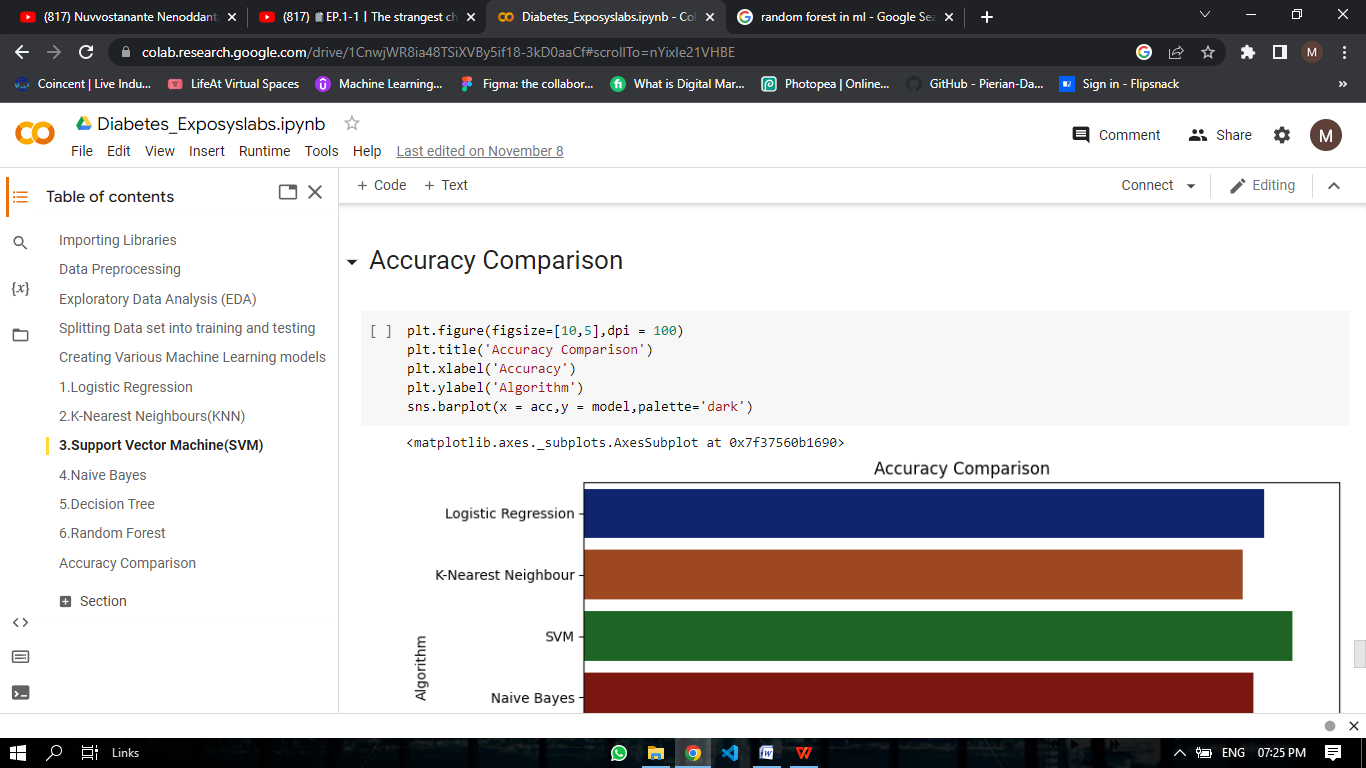
Random Forest:

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems in R and Python.

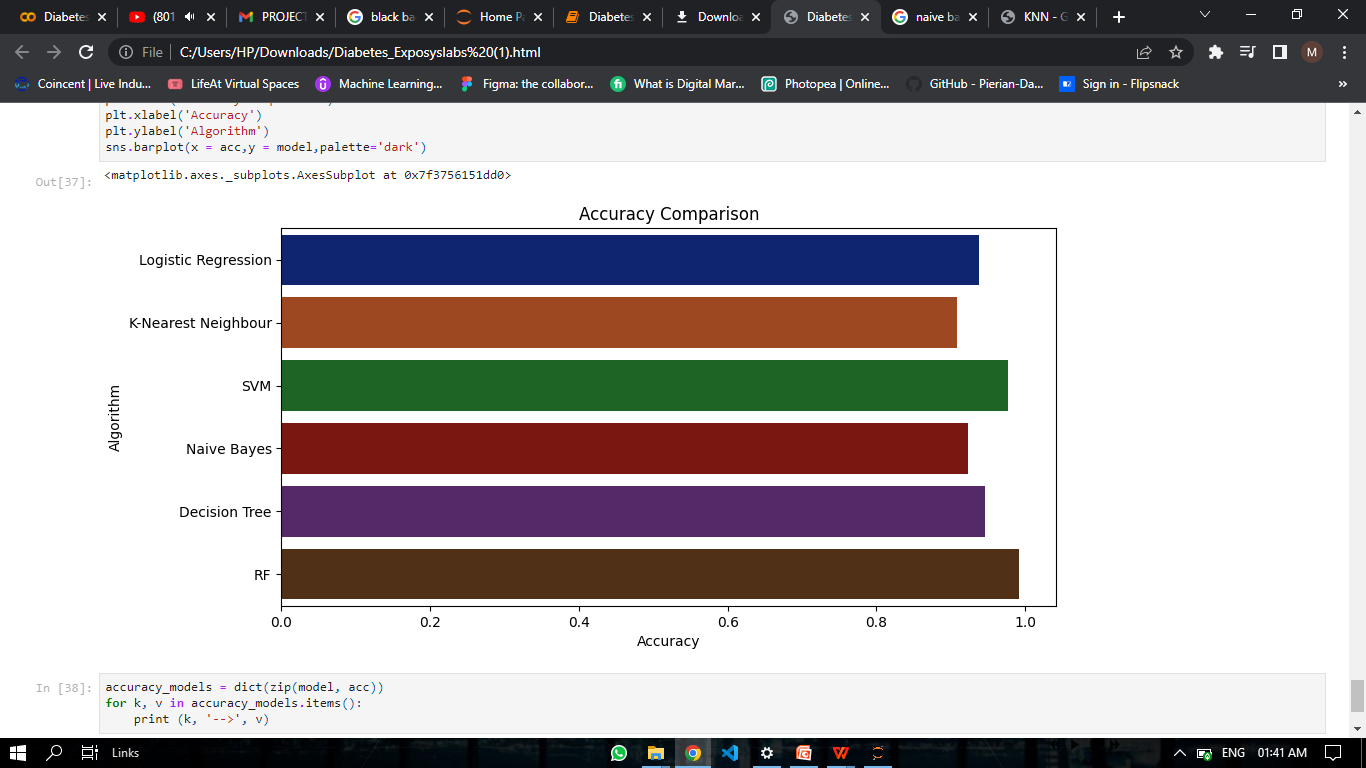


1. Comparing the Accuracies of the models

We compare the accuracies from all the models we have used to predict the best model for the task given.



Output:

****

From this we can conclude that “**Random forest**” has the highest accuracy of **99.23%**

Therefore we choose Random forest as the best model to give more accurate solutions for the early prediction of diabetes.

Link to the Google colab (Code): <https://colab.research.google.com/drive/1CnwjWR8ia48TSiXVBy5if18-3kD0aaCf#scrollTo=nYixIe21VHBE>

**Conclusion:**

Thus we can easily find the diabetes at an early stage reducing the seriousness of the disease. We used logistic regression, K-nearest neighbor, Support vector machine, Naïve Bayes, Decision tree and Random Forest with the accuracies of 0.938, 0.907, 0.769, 0.923, 0.946 and 0.992 respectively. Random forest was chosen to be the best model fit for this data and accurate predictions.