Likelihood Prediction of Early Stage Diabetes

Coursera Capstone

Background

- Diabetes is one of the fastest growing chronic life threatening diseases affecting millions of people every year
- The asymptomatic phase of the disease is relatively long and people suffering from diabetes remain undiagnosed.
- Hence early stage prediction of diabetes is important to start treatment and introduce lifestyle changes.

Problem

To predict the early stage risk of Diabetes based on common and less common signs and symptoms to aid in treatment of the disease.

Data

- This data has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.
- I have obtained this data from UCI Machine Learning repository.

Data Description

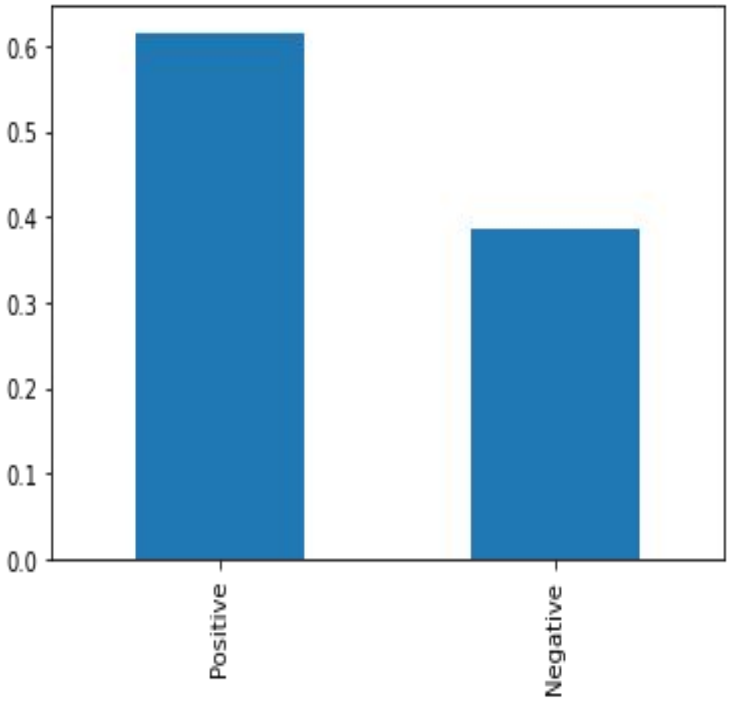
- The dataset contains 520 observations and 17 variables including the target variable.
- Some of the variables are Age, Gender, Polyuria, Polydipsia, etc.
- The target variable is denoted as “Class”.
- Most of the variables are categorical with “Yes” or “No”.
- Age is the only continuous variable.

Exploratory Data Analysis

To understand the characteristics of different variables in the dataset, I have performed univariate and bivariate analysis.

Calculation of Target Variable

There are 320 positive and 200 negative instances in the dataset. The division of “Class” in the dataset is shown in the bar chart.



Input variables - Categorical

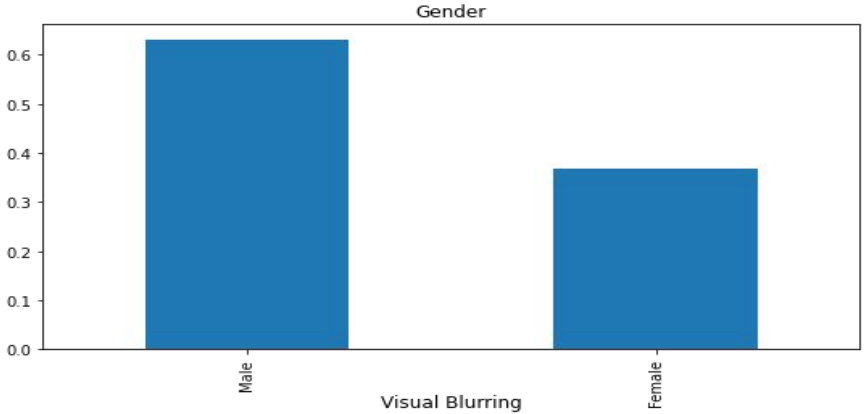
Few input variables which are easily identifiable looking at an individual like Obesity, Sudden Weight loss, Obesity etc have been analysed.

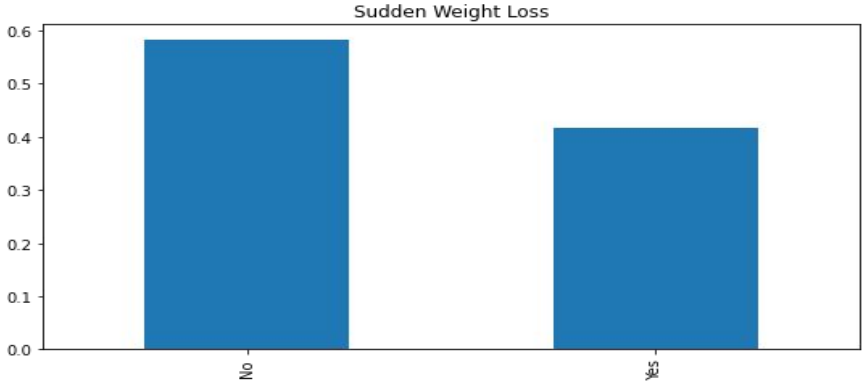
The graph is shown on the next slide.

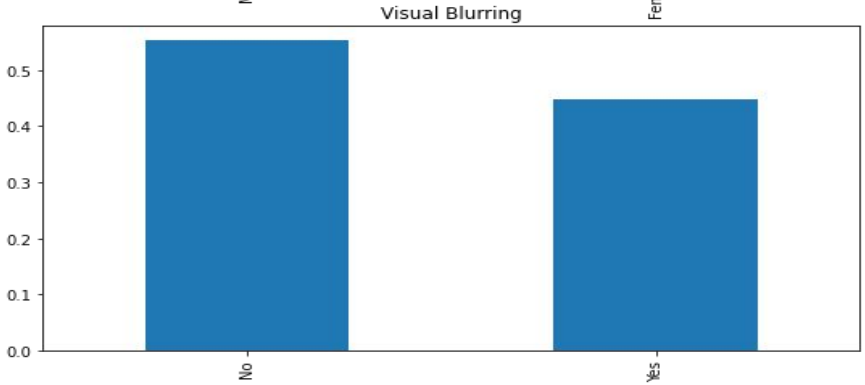
It can be inferred that:

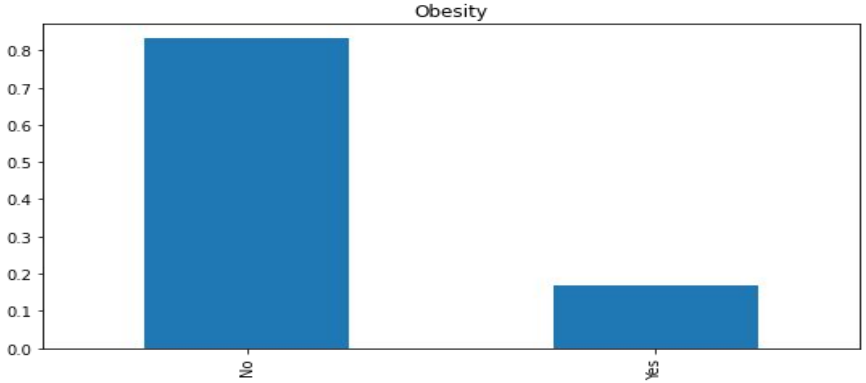
1. More than 60% of the individuals in the dataset are Males
2. The percentage of "No" in the categorical variables like Sudden Weight loss, Visual Blurring, Obesity is higher than the percentage of "Yes"

Input variables - Categorical





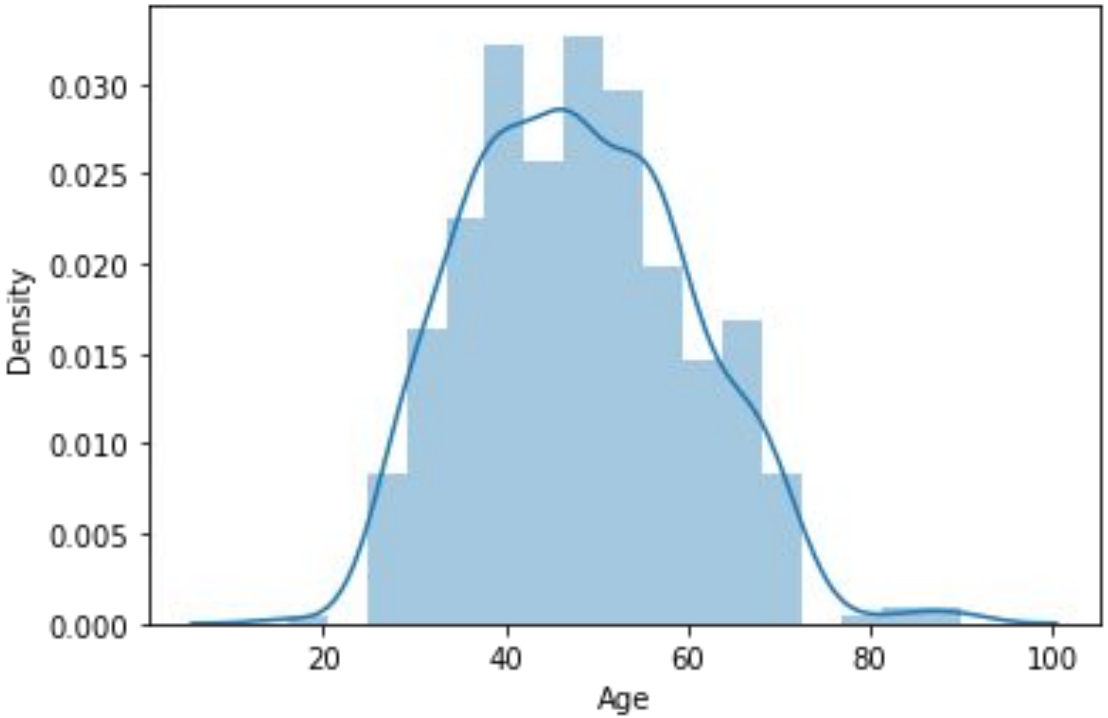




Input variables - Numerical

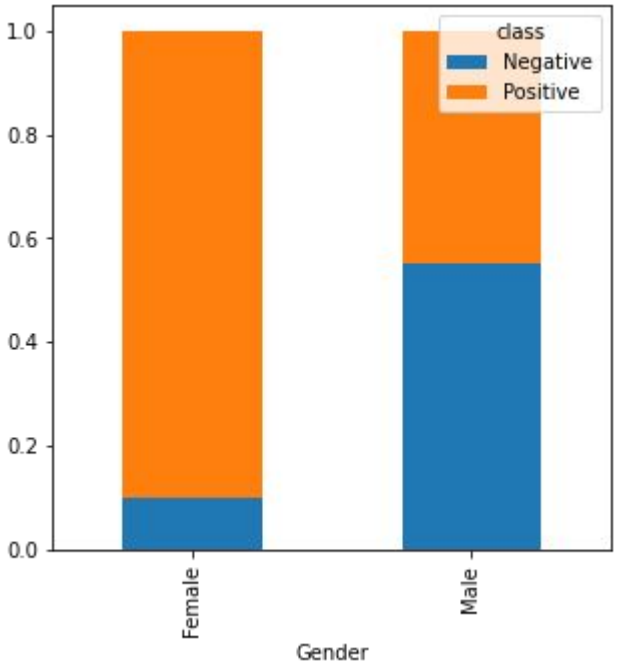
“Age” is the only numerical variable in this dataset. We need to check the distribution of data points for this variable so that it is not skewed to the left or right.

The “Age” variable is normally distributed which makes our analysis accurate and easier.



Relationship between Class and Gender

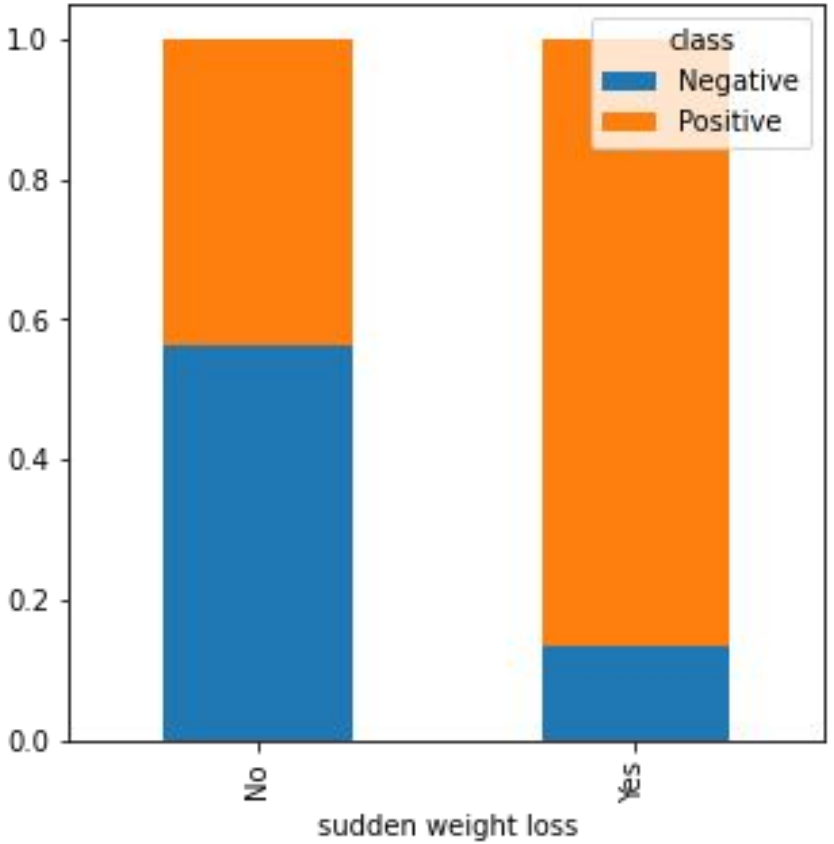
The graph above shows that the risk of diabetes is heavily dependent on gender. As seen, Females are at a higher risk than Males.



Relationship between Class and Sudden Weight loss

Sudden weight loss is one of the symptoms that is easily noticeable and is often thought to be related to diabetes

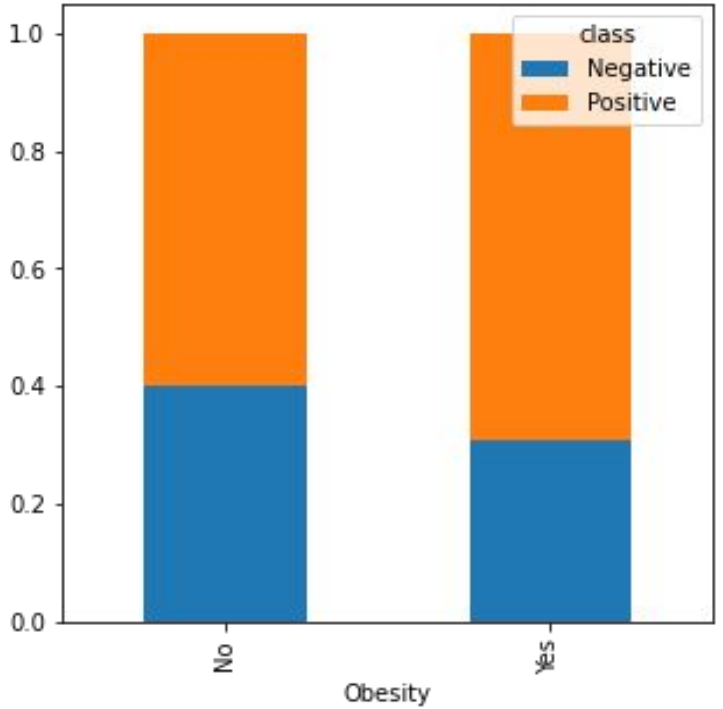
Sudden weight loss has a positive correlation with risk of diabetes. We can see from above that, "Yes" in Sudden Weight loss has more Positive cases than "No".



Relationship between Class and Obesity

It's a popular belief that obesity increases the risk of diabetes.

The "Yes" in Obesity has a higher risk of being Positive for diabetes. However, the difference between "Yes" and "No" for Obesity is not high for Positive.

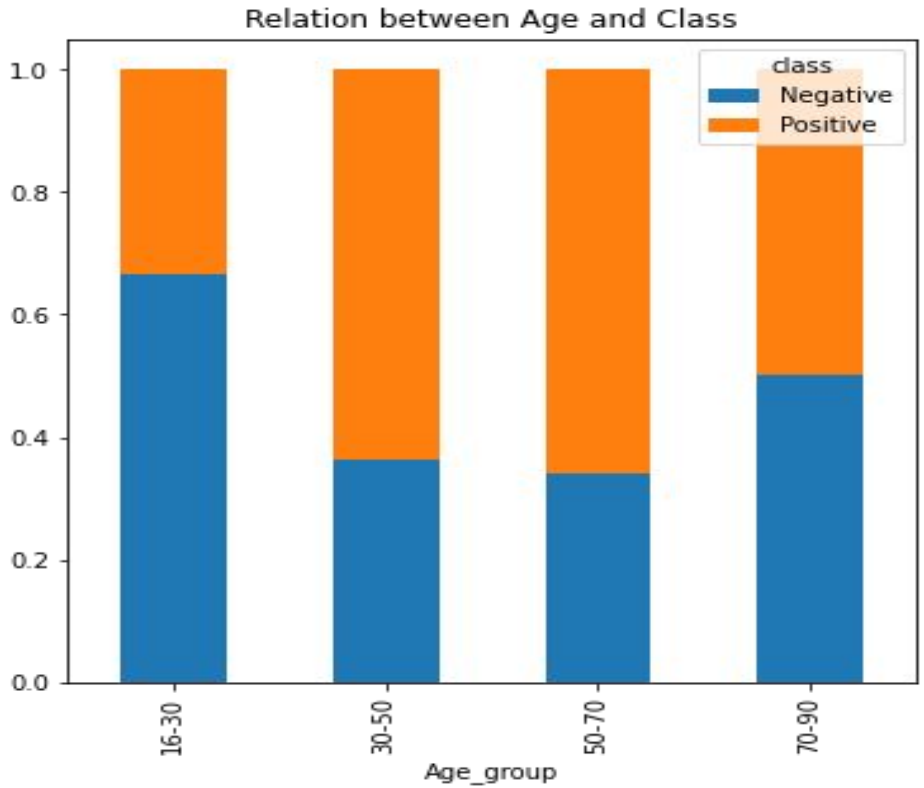


Relationship between Age and Class

Since age is a continuous variable, it has been divided into “Age groups” for the purpose of this analysis.

Age has been divided into 4 groups: “16-30”, “30-50”, “50-70”, “70-90”.

From the graph, it can be inferred that, the risk of diabetes is higher in age groups of 30-50 and 50-70.



Model Development and Evaluation

This is a classification since the output variable is “Class” which predicts whether an individual is at risk of diabetes or not.

The following algorithms have been chosen for modelling

1. Logistic Regression
2. Support Vector Machine (SVM)
3. K - nearest neighbors
4. Random Forest Classifier

Based on the evaluation metrics, the one with the **highest accuracy is Random Forest Classifier with an accuracy of 97.995%**, followed by Logistic Regression with an accuracy of 92.31%. Hence **Random Forest Classifier will be chosen as the final model** for this dataset.