Likelihood Prediction of Diabetes at Early Stage

1. Introduction

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. Early stage diagnosis of diabetes with the help of common and less common signs and symptoms is hence very necessary to initiate proper treatment and introduce lifestyle chances in the affected population.

As part of the Coursera Capstone project, I have taken a Classification problem to predict the early stage diabetes based on different symptoms of an individual. The data has been obtained from UCI Machine Learning Library.

2. Business 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.

3. 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.

3.1 Data description

The dataset has 520 observations and 17 variables including the target variable. Shown below is the information of the attributes.

Attribute Information:

Age 1.20-65

Sex 1. Male, 2.Female

Polyuria 1.Yes, 2.No.

Polydipsia 1.Yes, 2.No.

Sudden weight loss 1.Yes, 2.No.

Weakness 1.Yes, 2.No.

Polyphagia 1.Yes, 2.No.

Genital thrush 1.Yes, 2.No.

Visual blurring 1.Yes, 2.No.

Itching 1.Yes, 2.No.

Irritability 1.Yes, 2.No.

Delayed healing 1.Yes, 2.No.

Partial paresis 1.Yes, 2.No.

Muscle stiffness 1.Yes, 2.No.

Alopecia 1.Yes, 2.No.

Obesity 1.Yes, 2.No.

Class 1.Positive, 2.Negative.

As seen in the information above, most of the variables are categorical except the “Age” variable which is numerical.

The target variable is “Class” which can be Positive or Negative, indicating that a given individual can be at risk (or not) of diabetes based on different signs and symptoms described by other variables in the dataset.

3.2 Dealing with Categorical Variables

Since the categorical variables are in the form of “Yes” or “No” in the dataset, we need to convert them to numerical values. I have used sklearn’s LabelEncoder to do this.

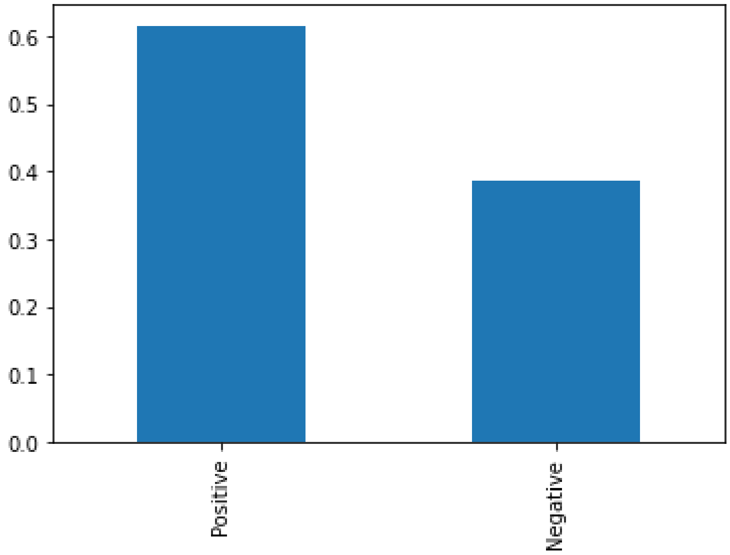
3.3. Exploratory Data Analysis

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

Calculation of Target Variable

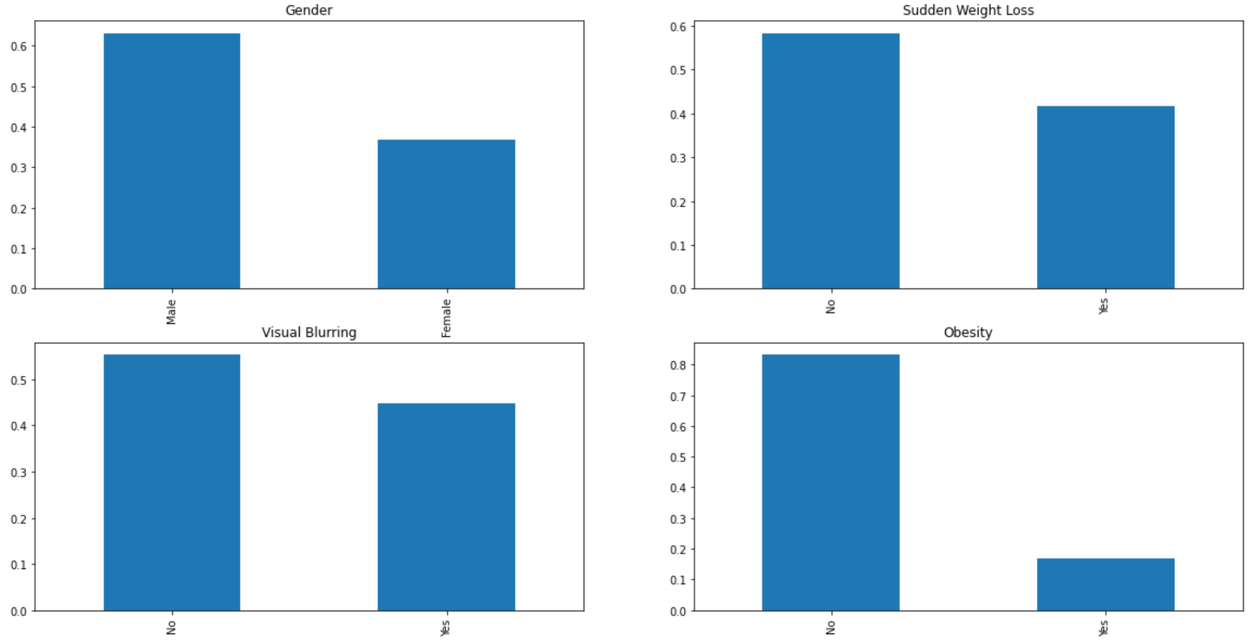
In this dataset, “Class” is the target variable which can be positive or negative - showing whether a specific individual can be at risk of diabetes based on signs and symptoms which are given as inputs.

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



Studying few input variables

The following graph shows the distribution of a few input variables in the dataset like: Gender, Sudden Weight loss, Visual Blurring and Obesity.



From the analysis we can infer the following:

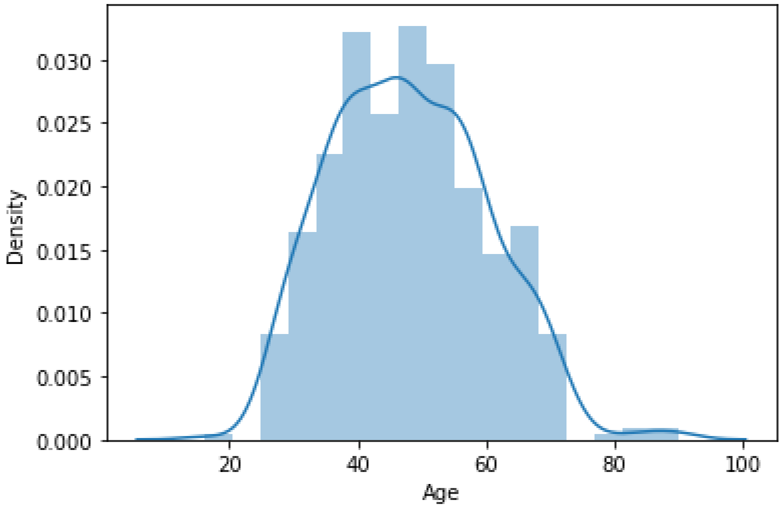
1. More than 60% of the individuals in the dataset are Males

2. The percentage of "No" in the categorical variables like Sudden Weight loss, Visual Blurring, Obesity is higher than the percentage of "Yes"

Understanding the Numerical Variable “Age”

“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.

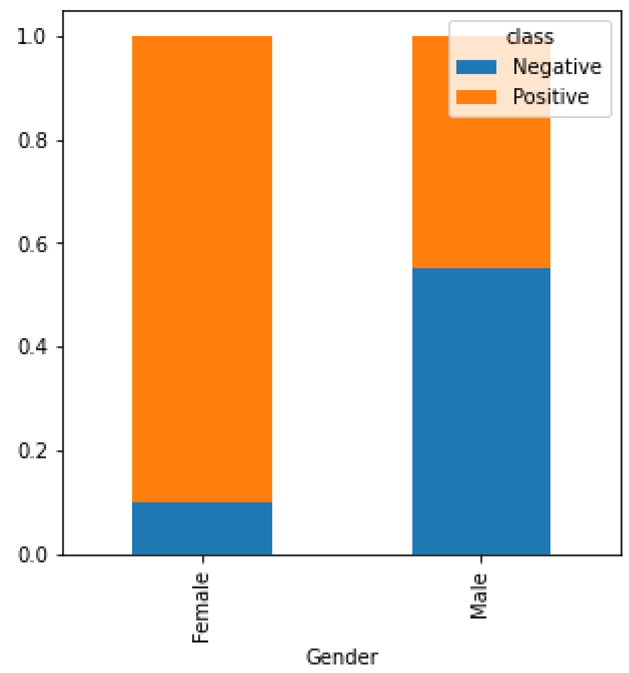
The graph below shows the distribution of “Age” data in the dataset.



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

Relationship between Gender and Class

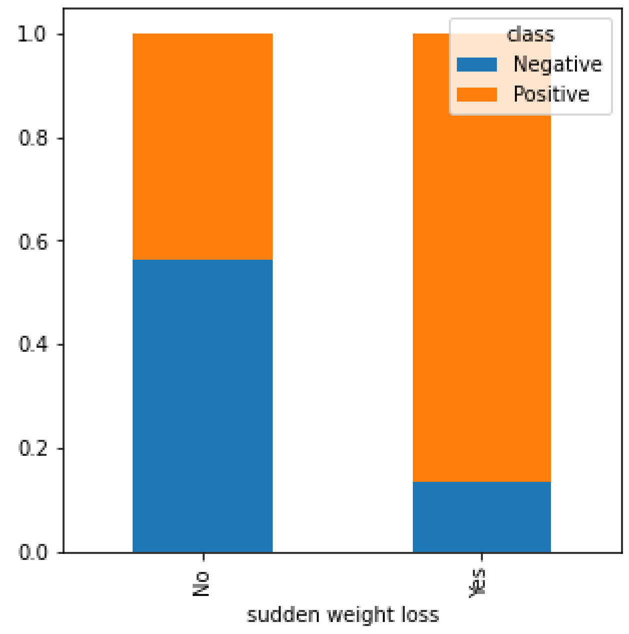
To understand the relationship between gender and the risk of diabetes, I have picked up these two variables.



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 Sudden Weight loss and Class

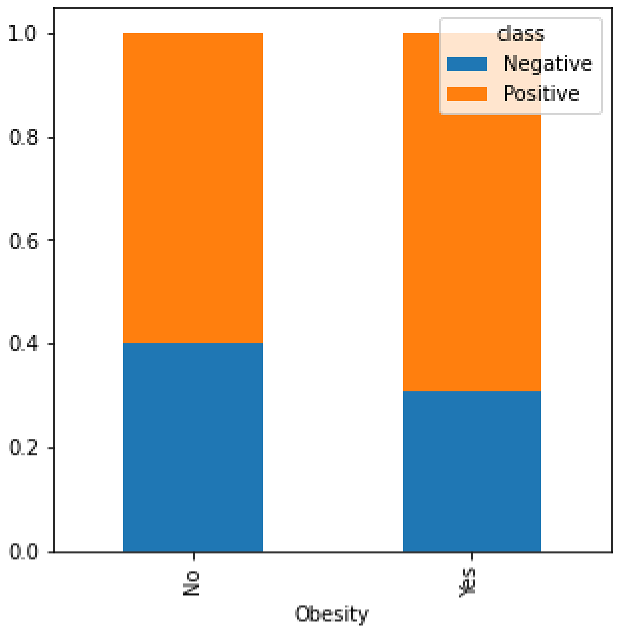
Sudden weight loss is one of the symptoms that is easily noticeable and is often thought to related to diabetes. Lets see if it holds a positive relationship with the risk of diabetes.



As seen above, 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 Obesity and Class

Its a popular belief that obesity increases the risk of diabetes. Lets see if this is true with the help of our dataset.



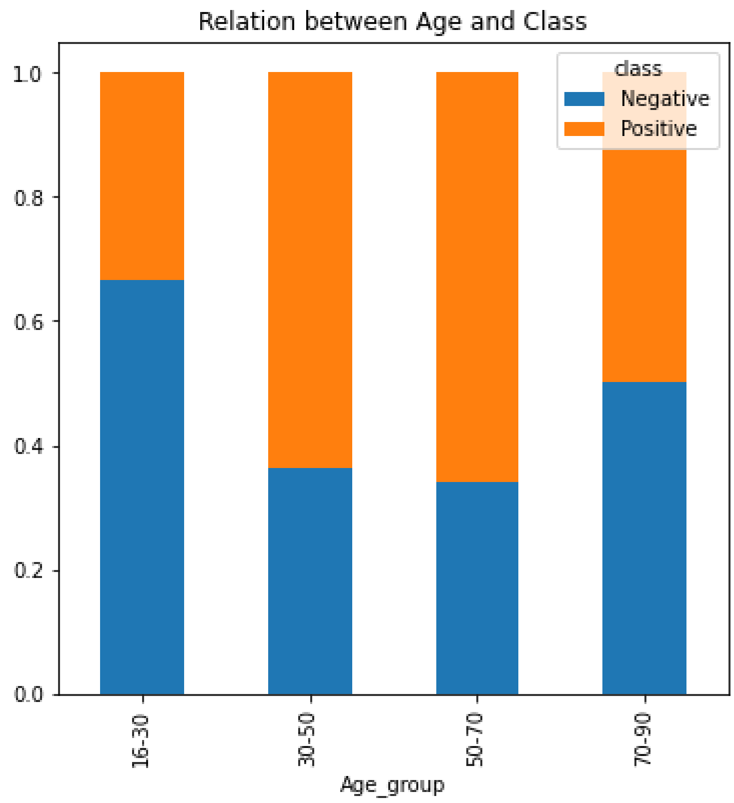
The "Yes" in Obesity has a higher risk of being Positive for diabetes. However, the difference between "Yes" and "No" for Obesity is not high for Positive.

Relationship between Age and Class

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

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

As can be seen from the graph below, the risk of diabetes is higher in age groups of 30-50 and 50-70.



4. Model Development

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

I have chosen four algorithms and developed the models.Finally the one with the better evaluation metrics will be chosen.

Algorithms chosen are the following:

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.

References:

Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125