Data Preparation, Cleaning and Feature Generation of a Diabetes Dataset

Seif ElAnsary†  
 CSCE Department  
 The American University in Cairo  
 ID:900221511  
 seif\_elansary@aucegypt.edu

MagdElDin AbdalRaaof  
 CSCE Department  
 The American University in Cairo  
 ID: 900211145  
 MagdElDin@aucegypt.edu

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

DATASET

In accordance with the previous report, we chose Alex Teboul’s diabetes dataset due to the fact that it is a large and a highly descriptive dataset. It contains 253,680 rows and 22 columns. This report aims to analyse each column, and the rationale behind the data processing of each one if it was processed.

ANALYSIS

1. Diabetes

The distribution of the Diabetes column is not uniform. A Diabetes score of zero means that the individual does not have diabetes. A Diabetes score of one entail that the individual is prediabetic and a score of two means that the person has diabetes.

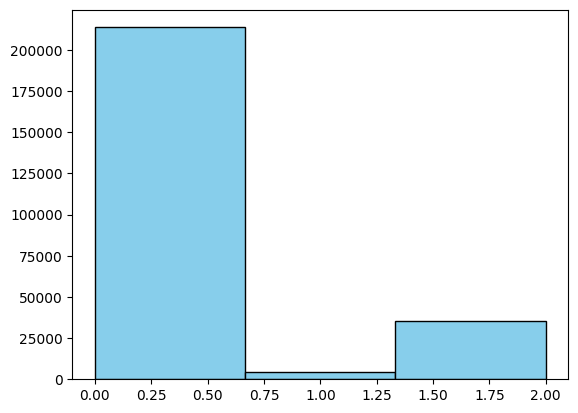


Figure 1

As shown in figure 1, around 80% of the people in the dataset are not diabetic. Even though the data is skewed a bit towards non-diabetic features, looking at the ratios and analysing columns with respect to each diabetes class will provide insightful data to train the model on.

1. High Blood Pressure

The HBP column is a binary classifier that labels people with blood pressure above a certain threshold as having high blood pressure (1), else they are classified as not having high blood pressure (0).

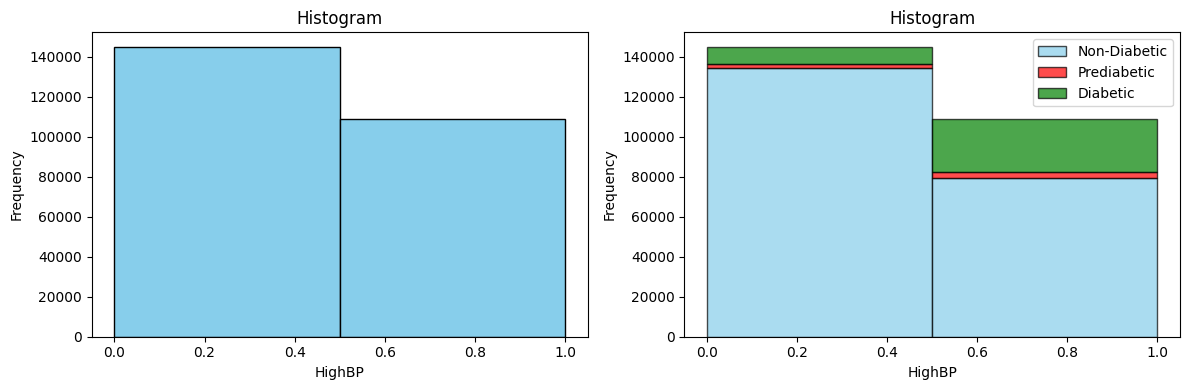


Figure 2

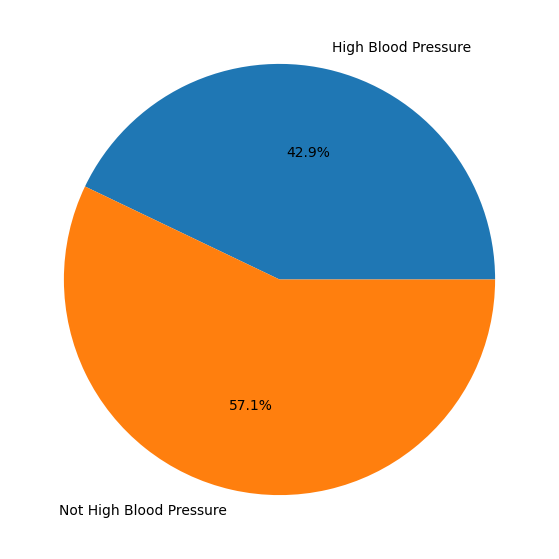


Figure 3

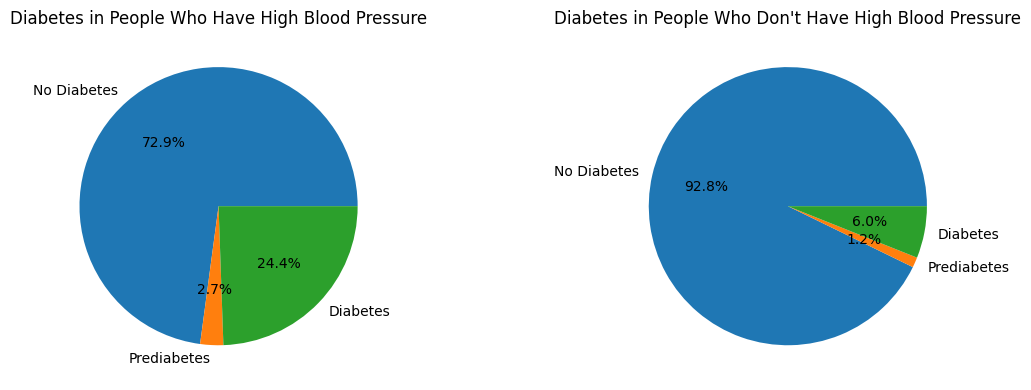


Figure 4

As shown by the pie charts, HPD is a significant indicator of whether a person has diabetes or not.

1. High Cholesterol

This column is also a binary classifier that is 0 if they individual does not have high cholesterol and is 1 if the individual’s cholesterol level is beyond a threshold.

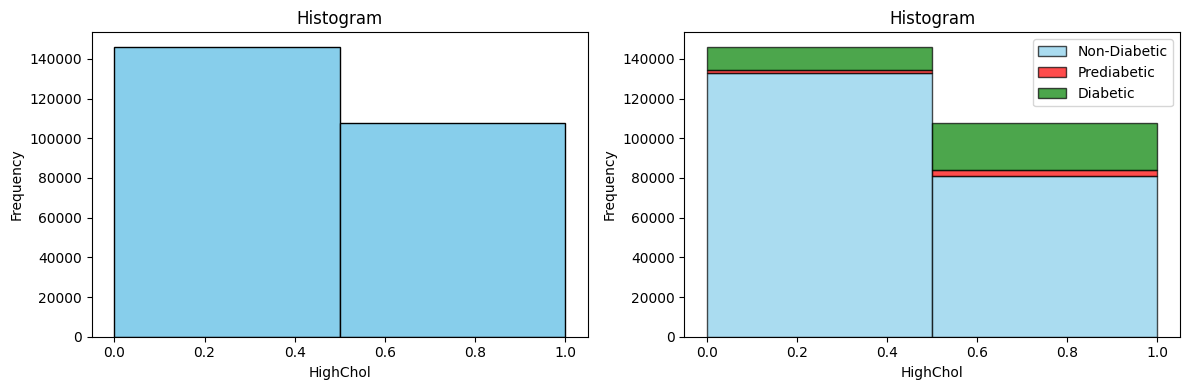


Figure 5

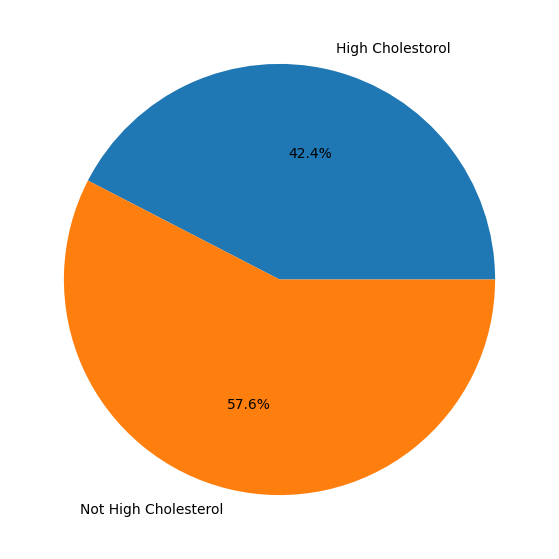


Figure 6

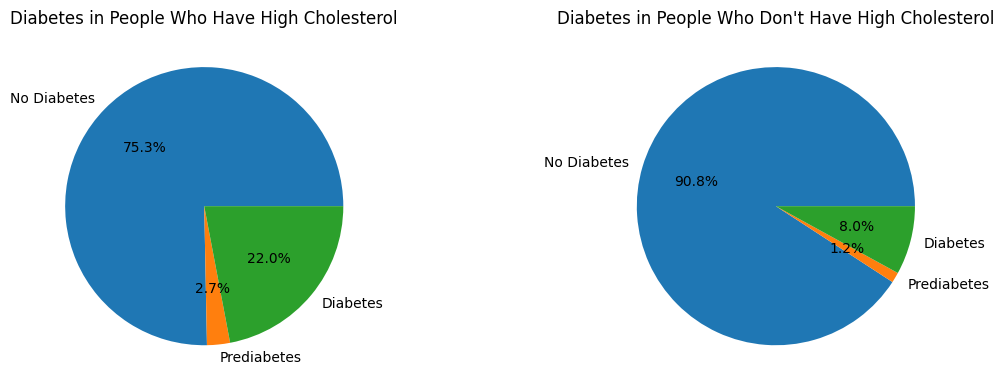


Figure 7

Figure 7 clearly shows that individuals which have high cholesterol levels increases the risks of diabetes and prediabetes.

1. **Cholesterol Check**

This column is a binary classifier that checks if a person checked their cholesterol levels during the last 5 years. Yes is 1 and no is 0.

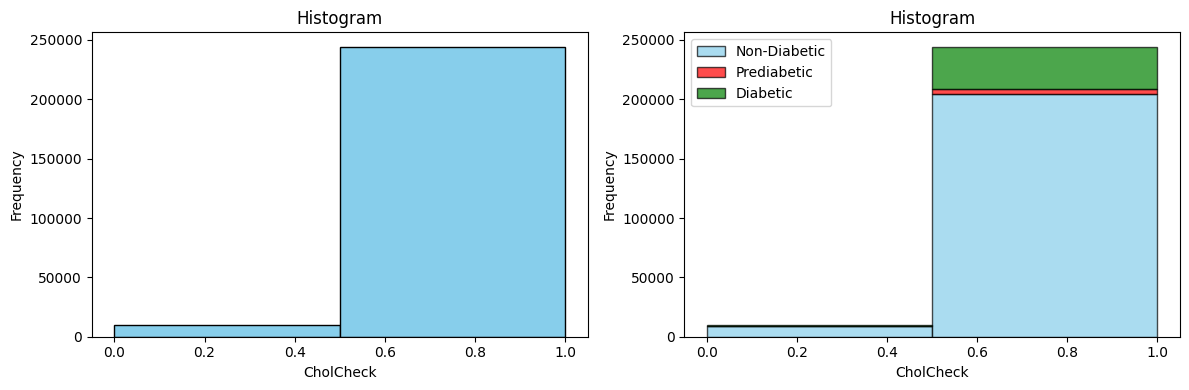


Figure 8

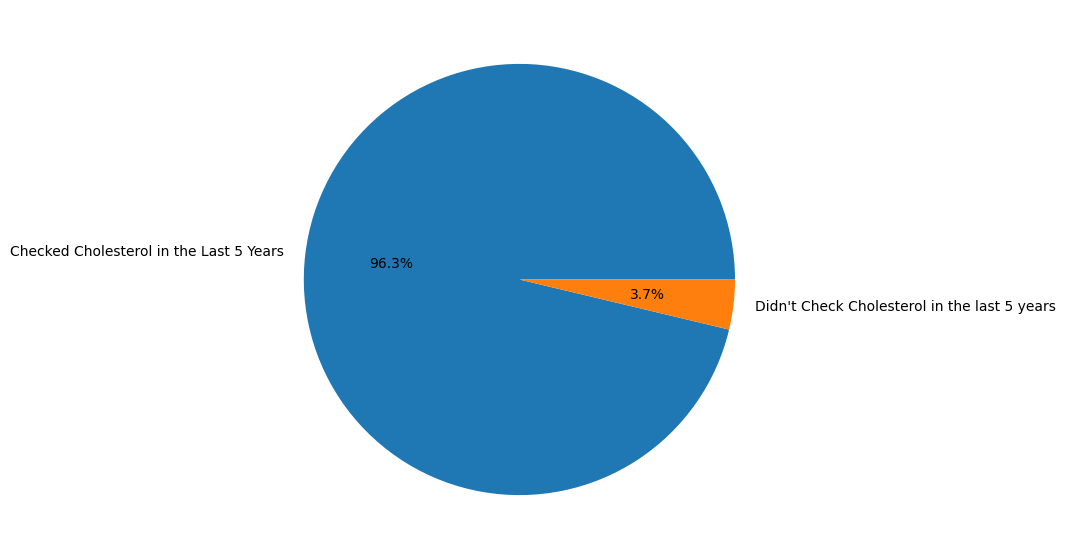


Figure 9

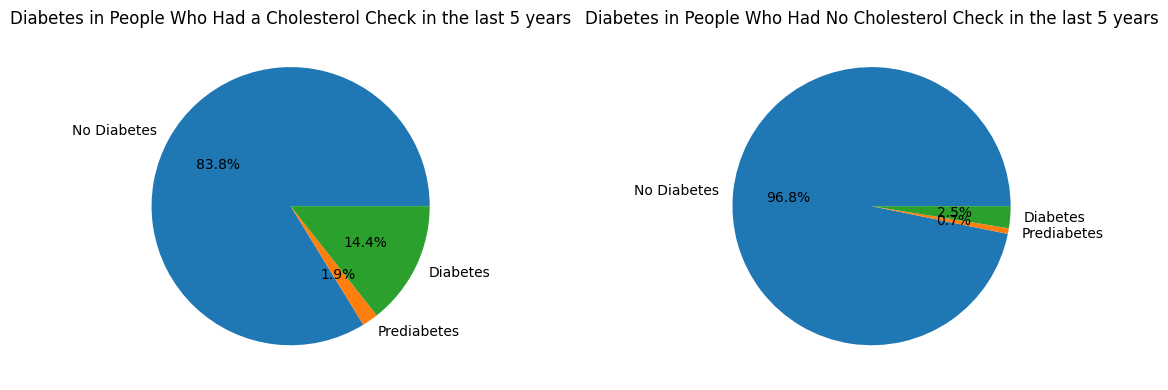


Figure 10

This feature might not seem that useful at first glance. It could imply that people getting their cholesterol checked are people who also suffer from various diseases that could lead them to check their levels more.

1. **Stroke**

Another binary classifier that checks if a person suffered a stroke before (1) or not (0).

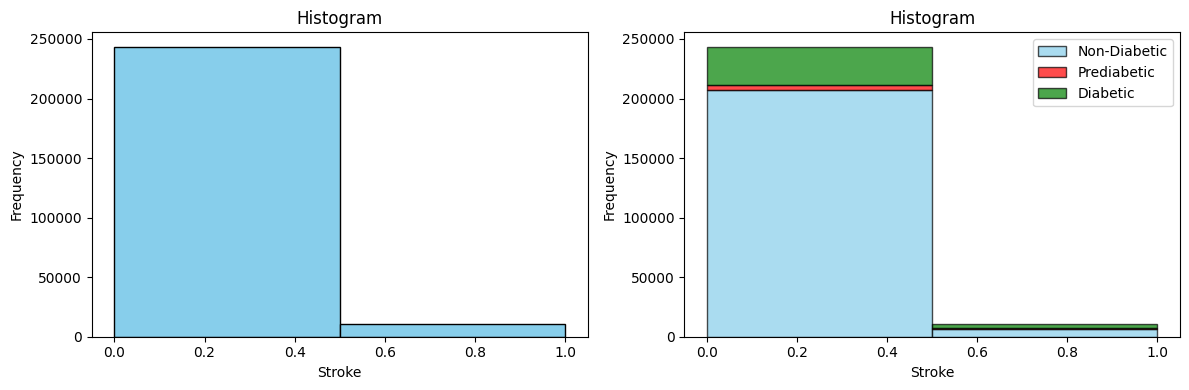


Figure 11

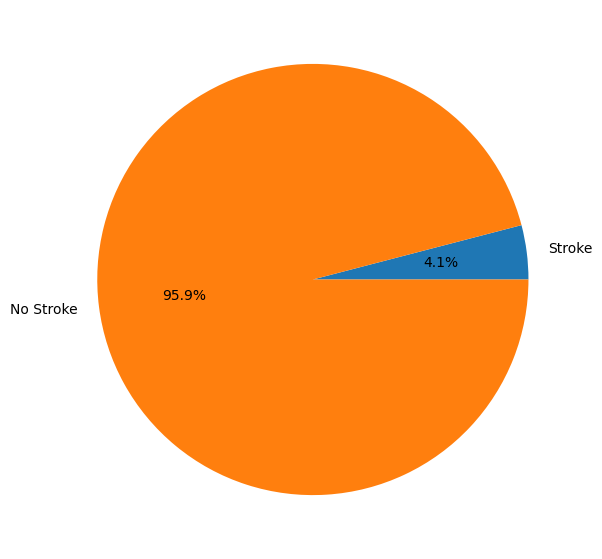


Figure 12

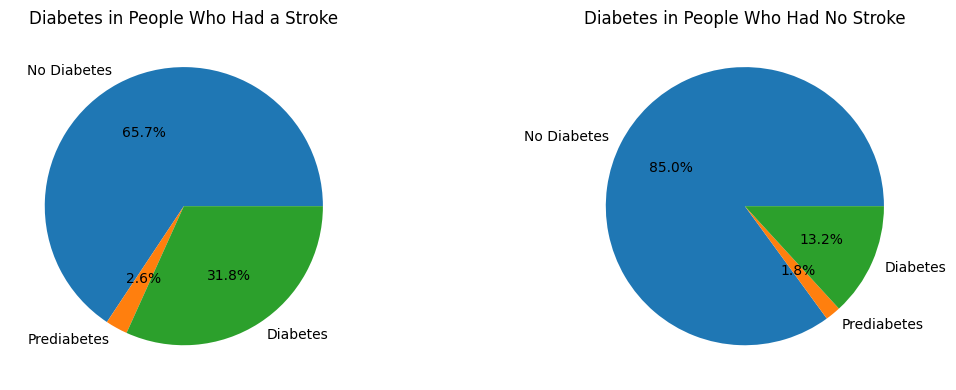


Figure 13

Expectedly, not many individuals have suffered a stroke. This is due to the fact that it is quite a rare event that does have a considerable mortality rate. However, those who had a stroke tend to have a higher probability of having diabetes than people who did not suffer from a stroke.

1. **Heart Disease or Attack**

This column is also a binary classifier that checks whether an individual had coronary heart disease (CHD) or myocardial infarction (MI) (1) or does not (0).

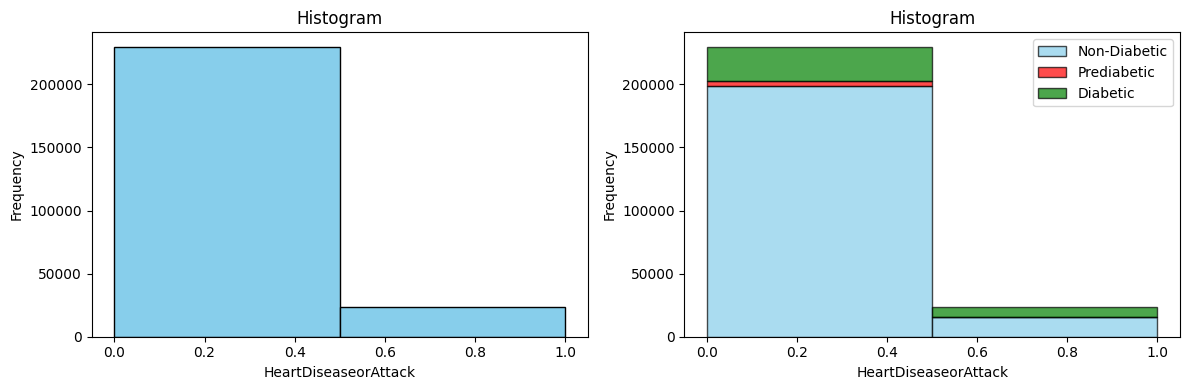


Figure 14

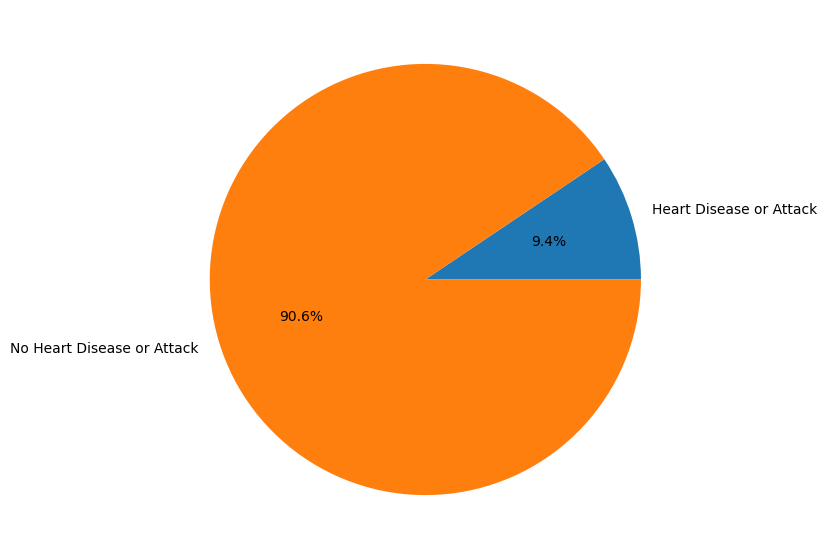


Figure 15

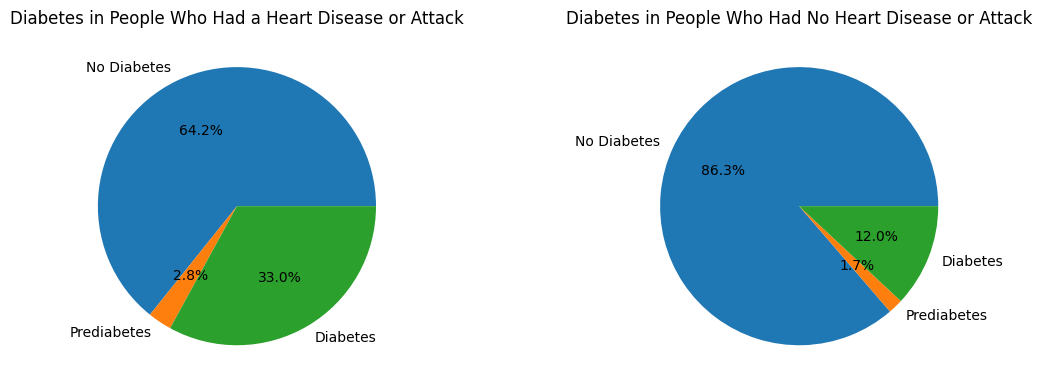


Figure 16

This is a good feature since it shows a correlation between heart disease or attack and having diabetes.

1. **Physical Activity in the Last 30 Days**

This column is a binary classifier that measures if a person did physical activity in the last 30 days. If yes, then it is 1, else 0.

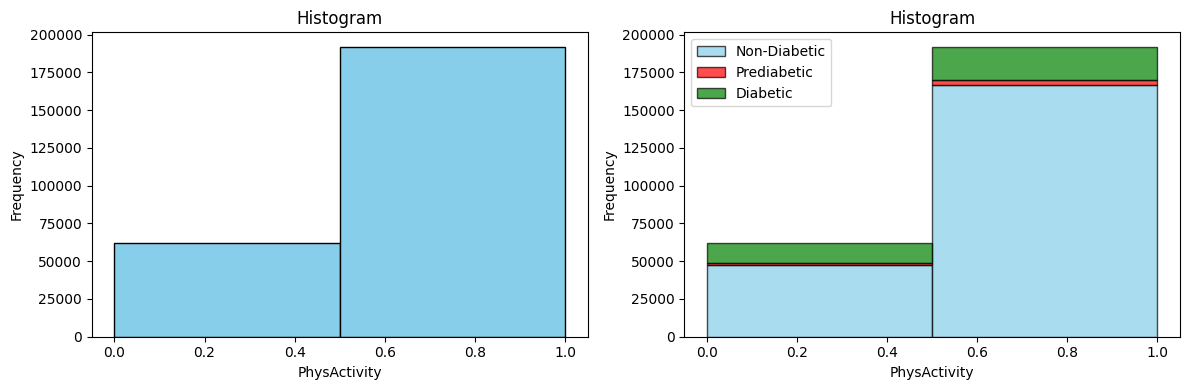


Figure 17

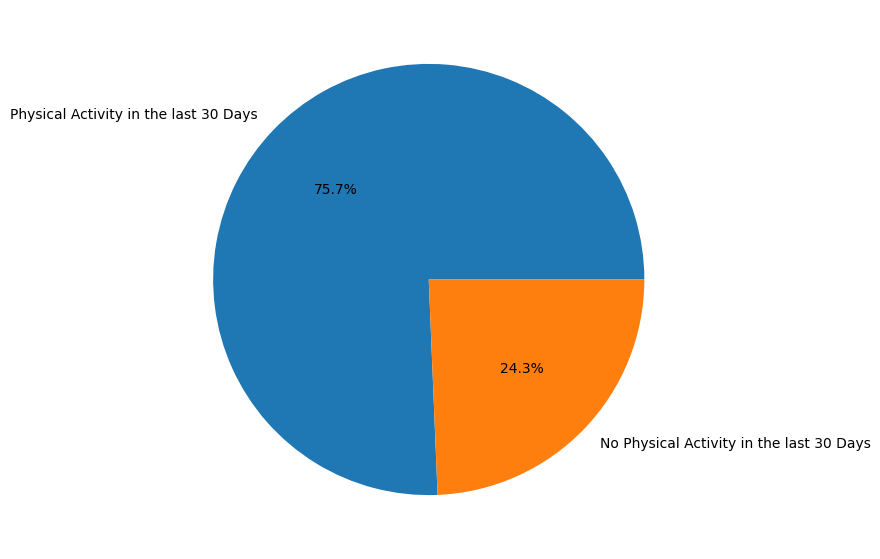


Figure 18

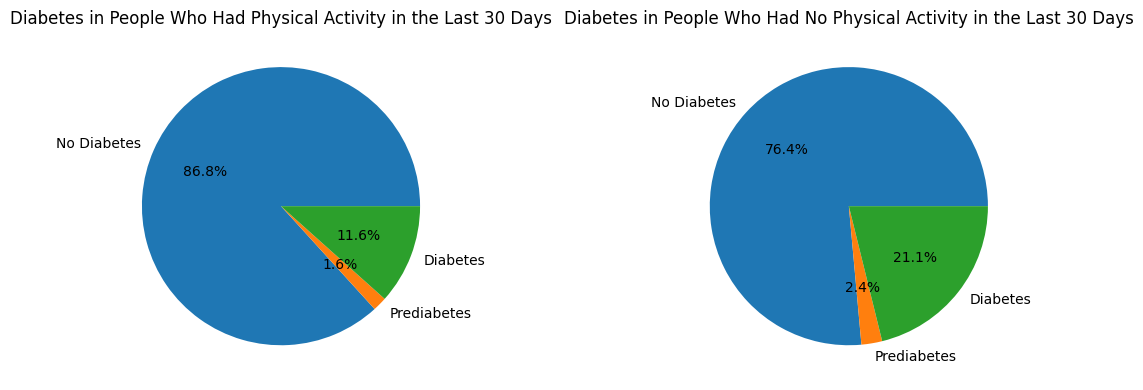


Figure 19

Intuitively, physical activity decreases the chance of a person having diabetes by around 10%. This is a pretty descriptive feature to have.

METHOD

DATASET REVIEW

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00