PIMA Indians Diabetes Prediction

What is diabetes ?

- Diabetes is a metabolic disease that causes high blood sugar. This is caused because of higher-than-normal blood sugar levels which is because of pancreas not secreting enough insulin to maintain normal blood sugar level.

- Undiagnosed diabetes can be fatal and can damage eyes, nerves, kidneys and other vital organs.

- There are two types of diabetes:

  - Type 1 diabetes : this type of diabetes is an autoimmune disease. The immune system destroys cells in the pancreas that disturbs the regular insulin generation cycle.

  - Type 2 diabetes : This type of diabetes is when the body becomes resistant to insulin, It also does not accept insulin injected externally, which causes sugar build up in blood.

Why we chose Diabetes as our project topic ?

- We are intrigued by the fact that close 420 million worldwide suffer with diabetes.

- Given the fact that diabetes can be dangerous if not diagnosed at an early stage. With proper treatment the patient can continue to live life normally with minor dietary modifications.

- Another motivation for us to choose diabetes is because we know of people among friends and family that have diabetes.

- This made us wonder what we could do to improve the current process of diabetes prediction.

- We decided that using machine learning will improve the accuracy with which we can predict if a patient has diabetes or not based on different factors.

#Facts

- About 422 million people worldwide have diabetes.
- Diabetes is one of the leading causes of death worldwide
- Type 2 diabetes is much more common than type 1 diabetes.
- Less than 50 per cent of all Canadians can identify less than half of the early warning signs of diabetes.
- Only 33 per cent of Canadians are aware that stroke is a complication of diabetes.
- Only 40 per cent of Canadians identified heart disease is a complication of diabetes.

Objective:

- Our main goal is to accurately predict if a patient has diabetes or not based on several different factors which include blood glucose level, blood pressure, skin thickness, insulin , BMI , age and number of pregnancies.

- The results from this project can be used on wearable devices that monitor some of the important parameters with which we can detect diabetes such as blood oxygen meter which was recently introduced in apple watch series 6.

Data Source

Kaggle Data source link: https://www.kaggle.com/uciml/pima-indians-diabetes-database

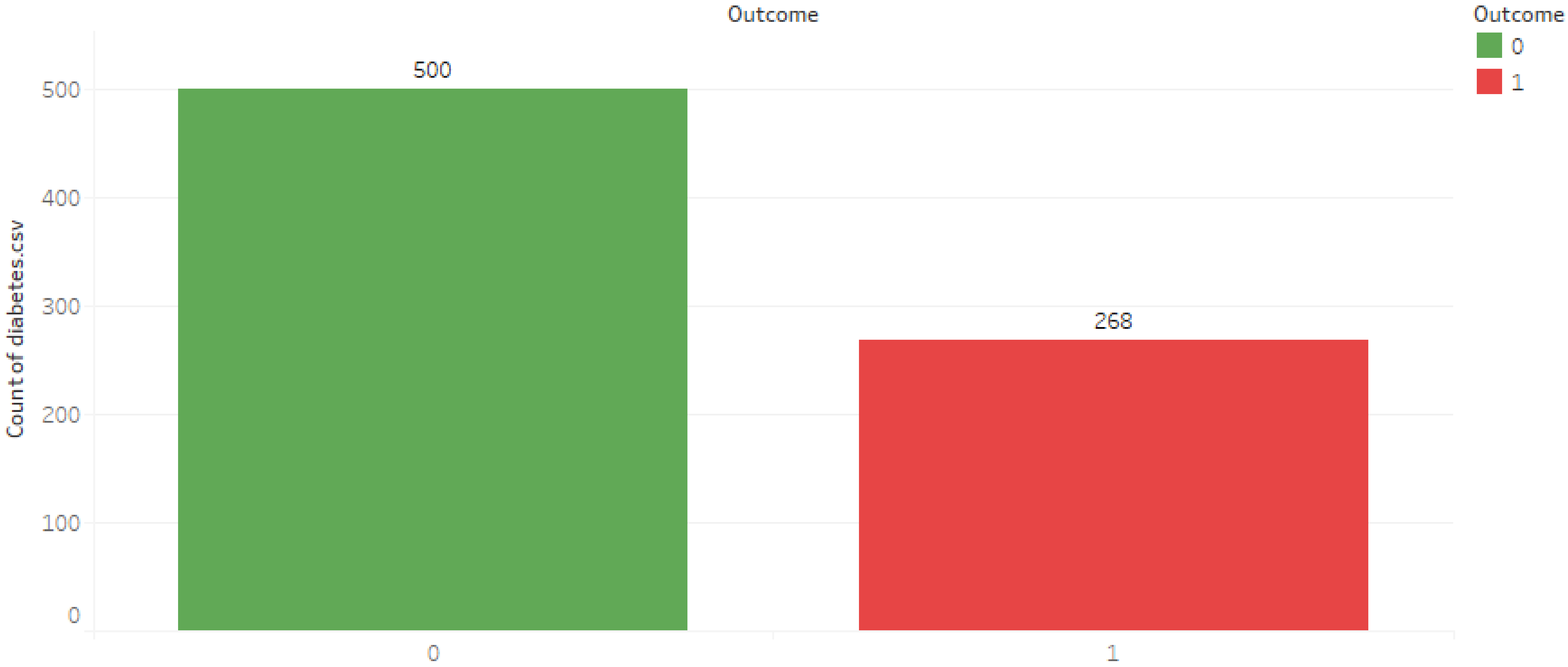
Dataset Description

Description: The dataset consists of 9 variables and 768 patient records. Following are the variables:

Pregnancies: Number of pregnancies.
Glucose: Plasma glucose concentration in an oral glucose tolerance test.
Blood Pressure: It can be defined as the pressure of blood circulation against the walls of blood vessels. It is measure in millimeters of mercury (mmHg).
Skin Thickness: Triceps skin fold thickness.
Insulin: It helps regulate blood sugar levels in the body.
BMI (Body mass index) : It measures body fat based on height and weight .
Age: Age of patients (in years)
Outcome: (target variable) Class 0 indicates that patient is non-diabetic and class 1 represents patient is diabetic.

Overview

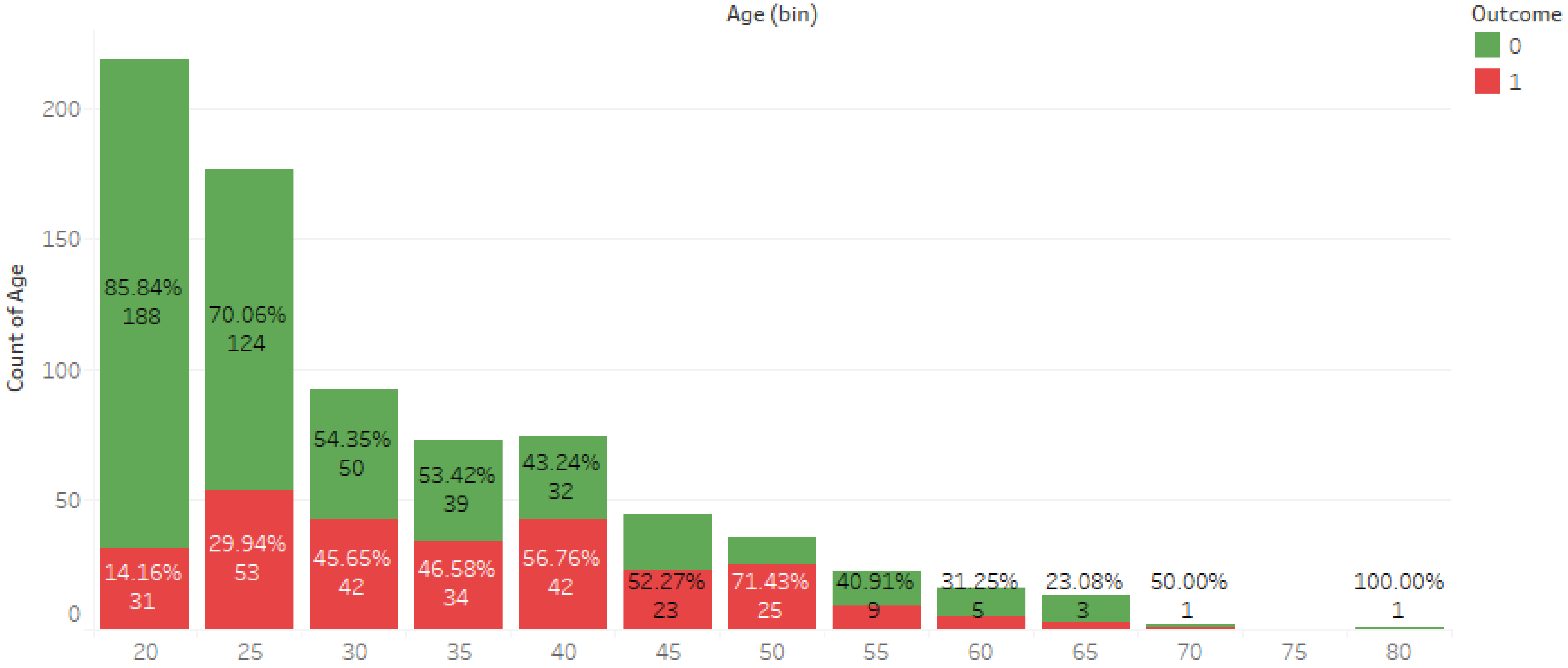
Overview



From the above chart we can see that out of 768 records there are 500 non diabetic and 268 diabetic with labels 0 and 1 respectively

Increase in number of diabetes patients with age

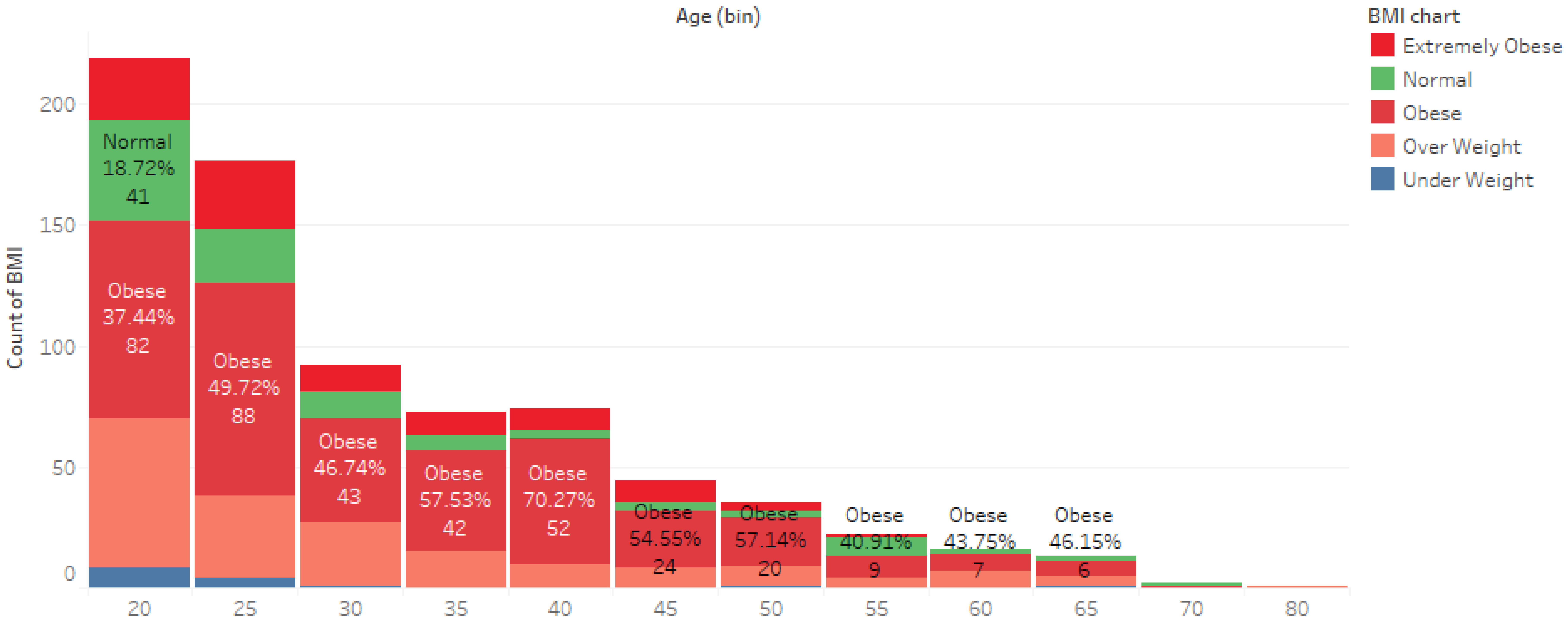
Increase in number of diabetes patients with age



From the above chart, we can see that with increase in age the number of diabetic patients in each age bin are increasing, which is a good indicator that older people are more at risk of diabetes. Furthermore, 50-55 age range has the highest number of diabetic patients.

BMI distribution

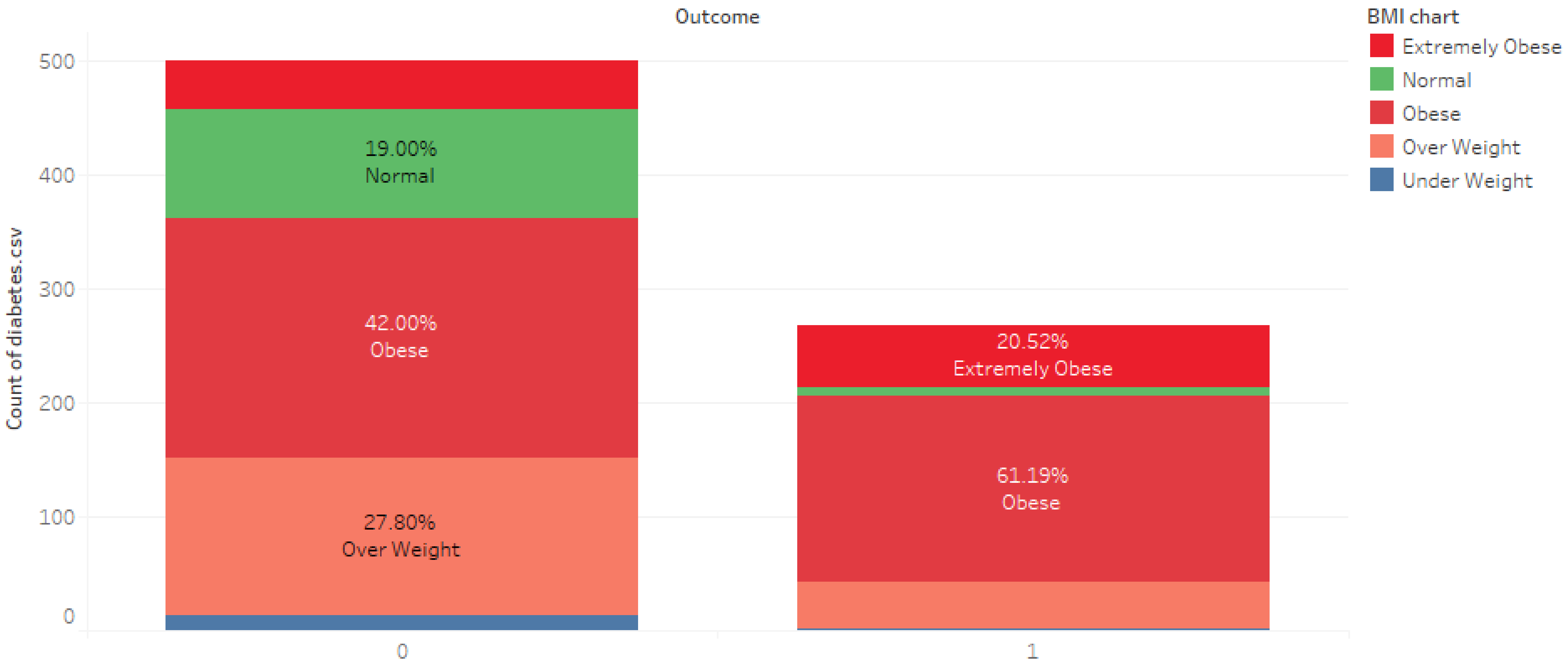
BMI distribution



From the chart, we can see that majority of obese are in the 40-45 age range and majority of extremely obese patients are in the 45-50 age range.

Relation between obesity and diabetes

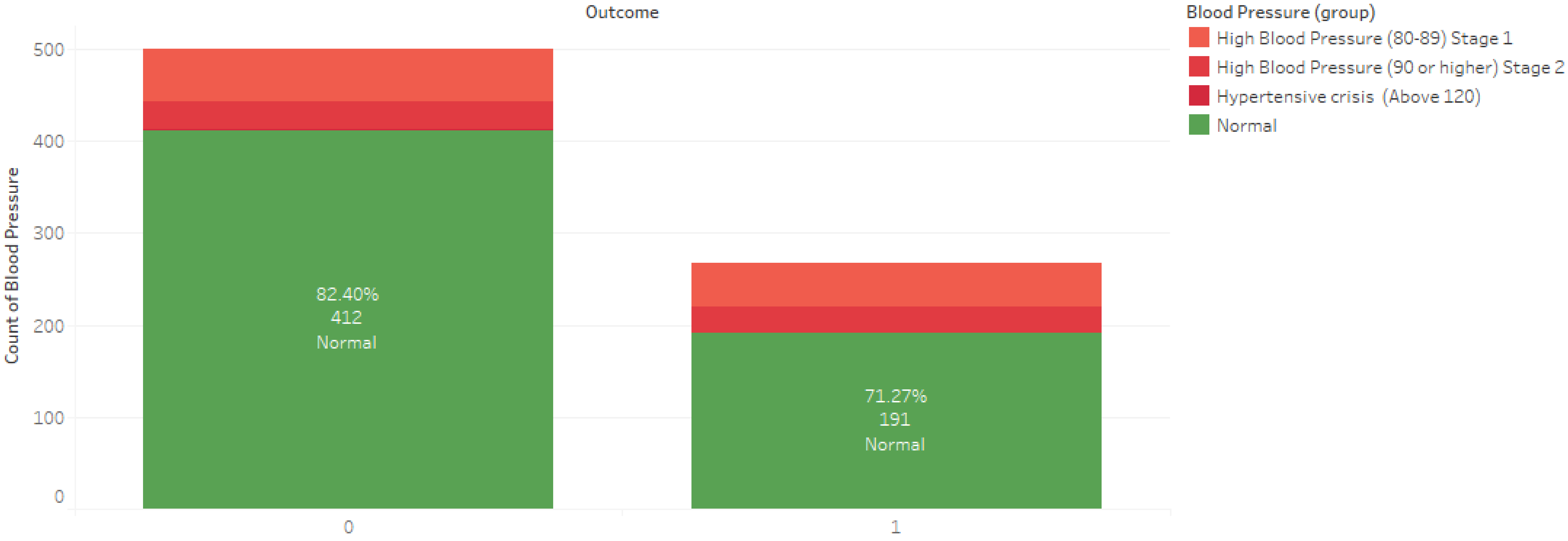
Relation between obesity and diabetes



From the chart, we can see that patients with obesity that have diabetes is higher than the percentage of patients that do not. Furhtermore, the percentage of patients that are extremely obese is higher in the patients with diabetes.

Relation between blood pressure and diabetes

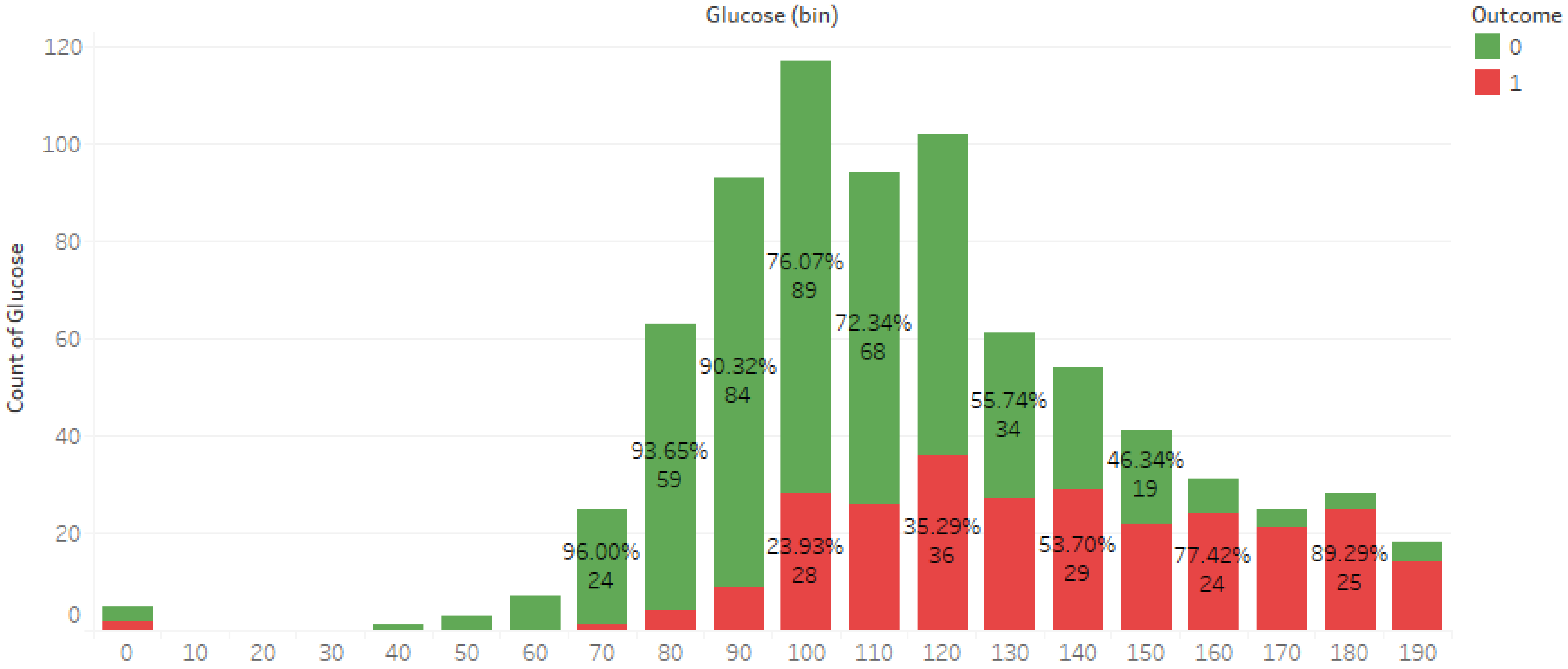
Relation between blood pressure and diabetes



From the chart, we can see that the number of patients with High blood pressure stage 1 and stage 2 is higher among the patients with diabetes. This is an indication that patients with increased blood pressure are at a higher risk of diabetes.

Relation between glucose and diabetes

Relation between glucose and diabetes

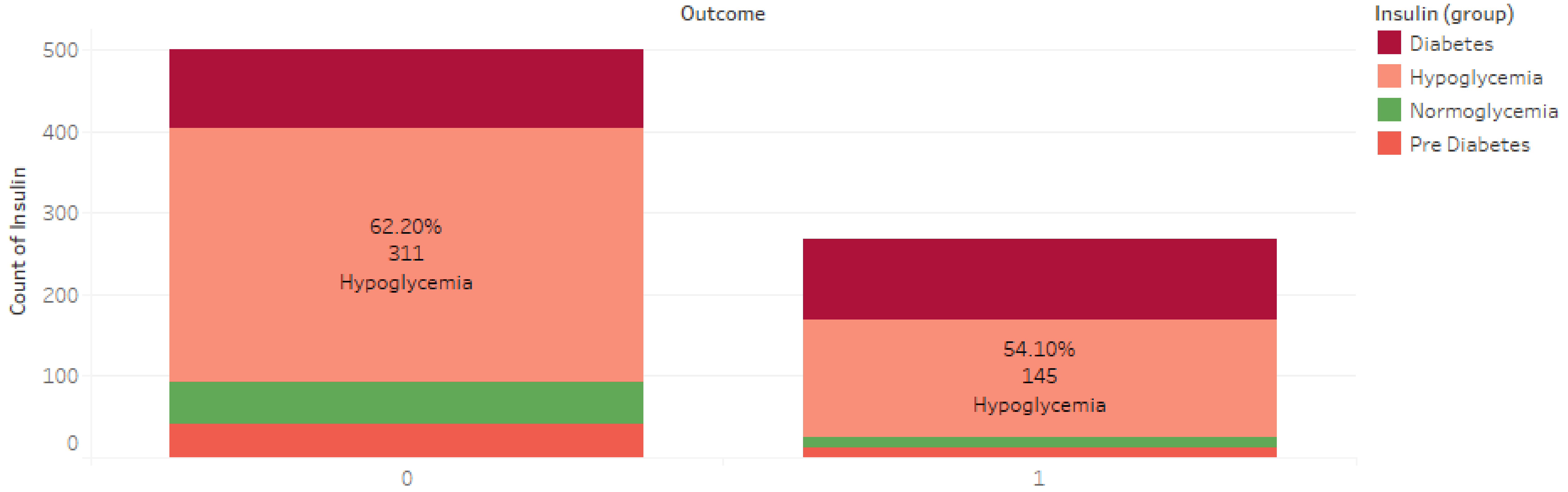


Blood glucose is the sugar that bloodstream transports to all cells in the body to supply energy. It is critical to maintain safe glucose levels to lower the risk of diabetes. It is measured by monitoring the level of sugar being transported at any given time.

From the chart, we can see that patients with glucose level of higher than 140 are at 50% risk of being diabetic.

Relation between insulin and diabetes

Relation between insulin and diabetes



Insulin is a hormone and it regulates metabolism of carbohydrates, fats and protiens by absorbing glucose from blood into liver, fat and muscles cells.

Hypoglycemia is when the insulin level is les than 70 which is lower than normal level and is dangerous. From the chart we can see that majority of the patients fall in this category.

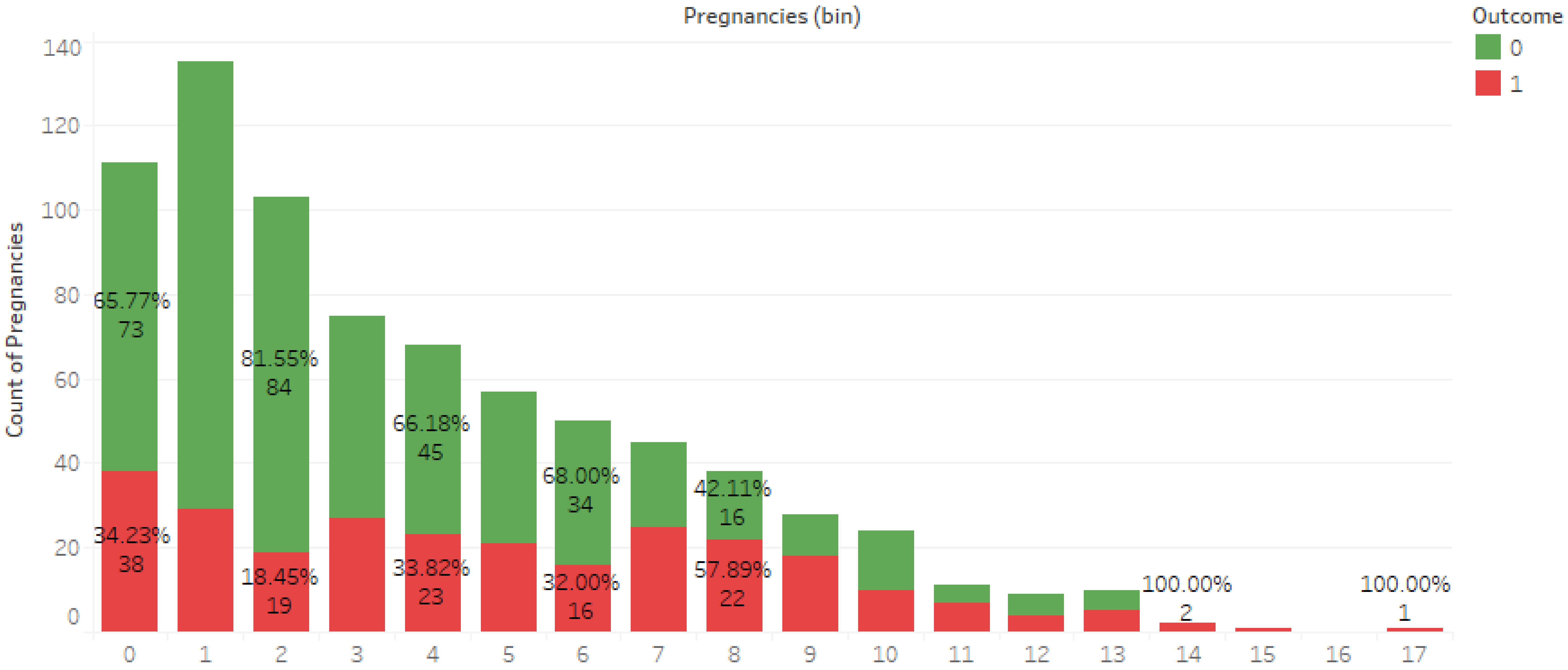
Normoglycemia is when the insuling levels are greater than 70 and les than 99 and from the chart we can see that very few patients fall in this category.

Pre Diabetes is when the insulin range in greater than 99 and less than 125, and they make a minor percentage in the dataset.

Diabetes us when the insulin level is greater than 125 and as we can see from the chart that there a sizeable number of patients that fall in this category

Relation between number of pregnancies and diabetes

Relation between number of pregnancies and diabetes

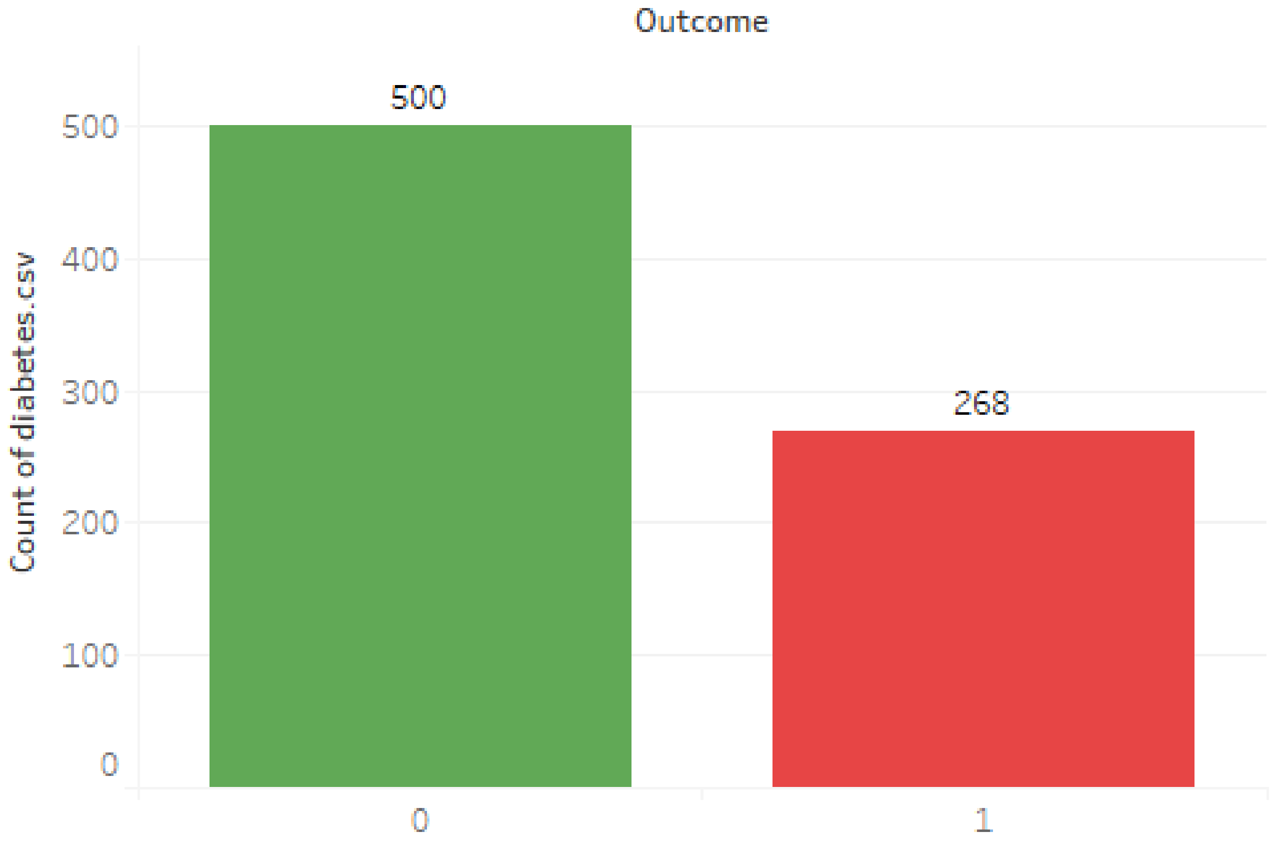


From the chart, we can see that higher the number of pregnancies then higher is the risk of being diabetic.

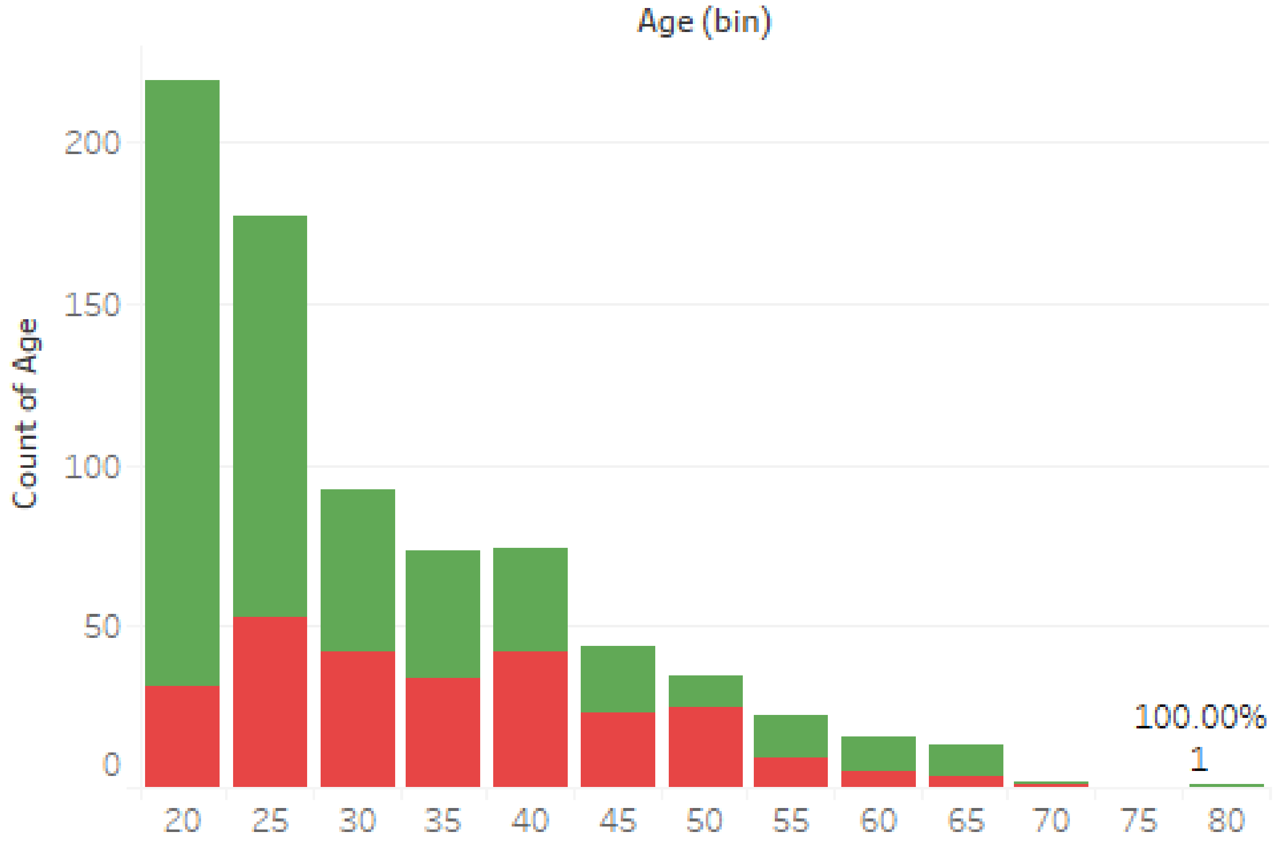
Pima indian diabetes (EDA part 1)

Pima indian diabetes (EDA part 1)

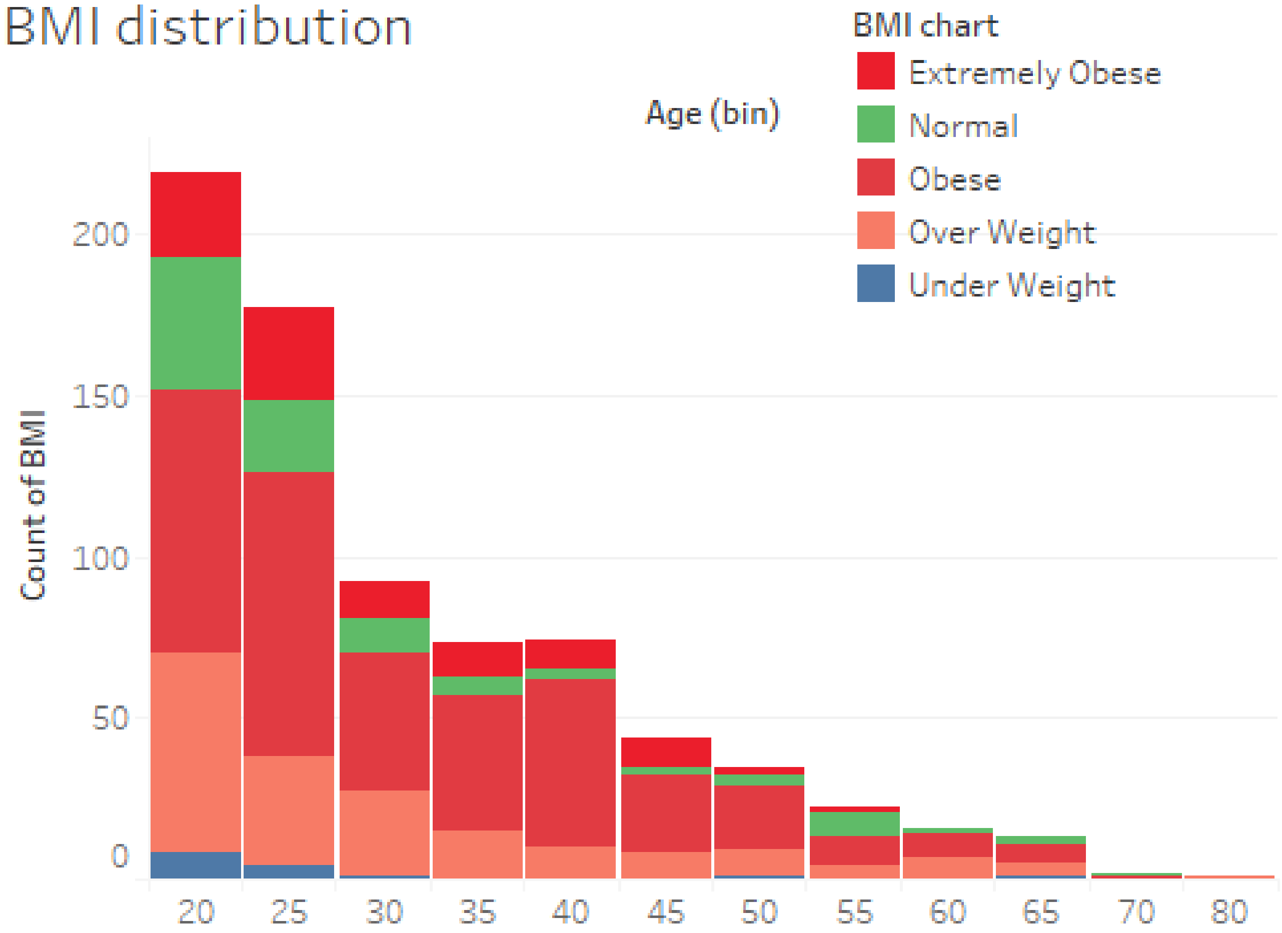
Age Distribution



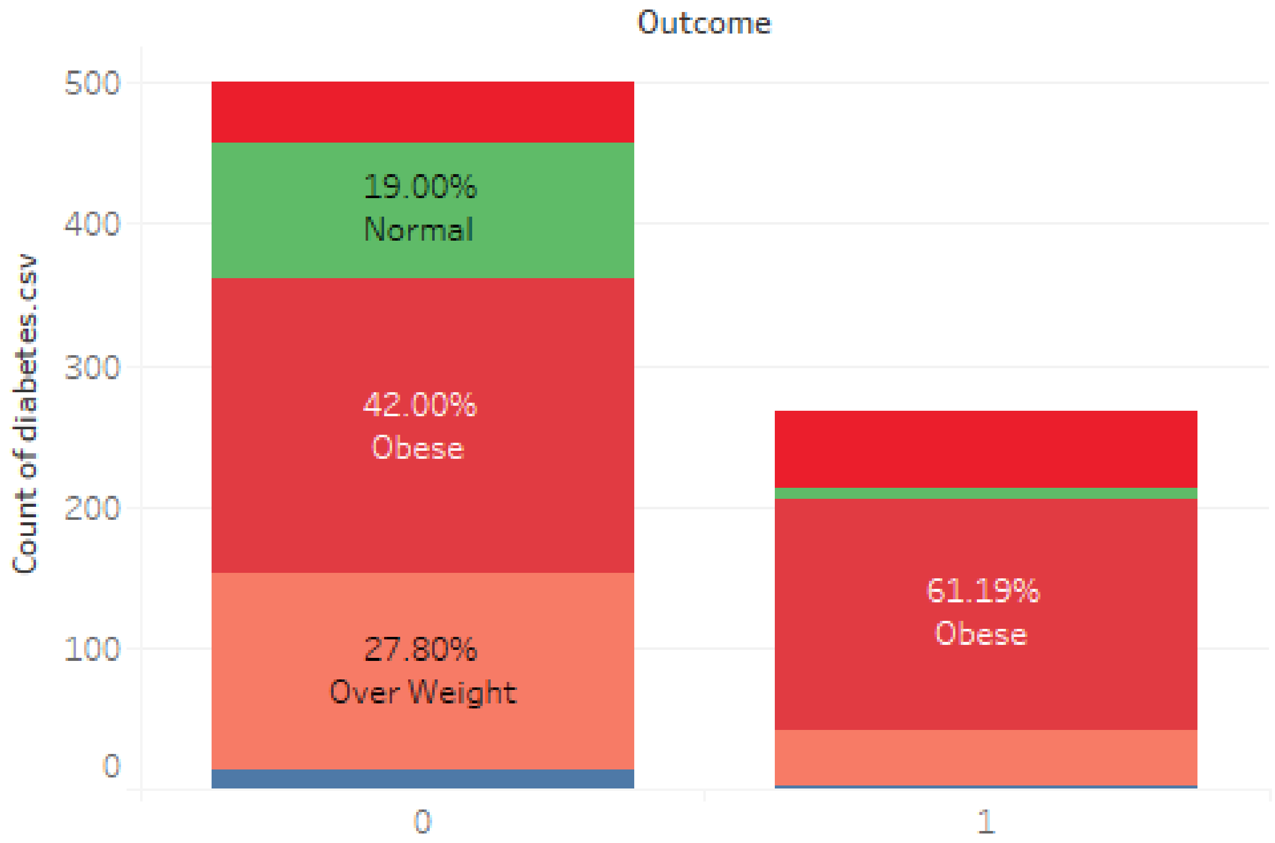
Increase in number of diabetes patients with age



BMI distribution



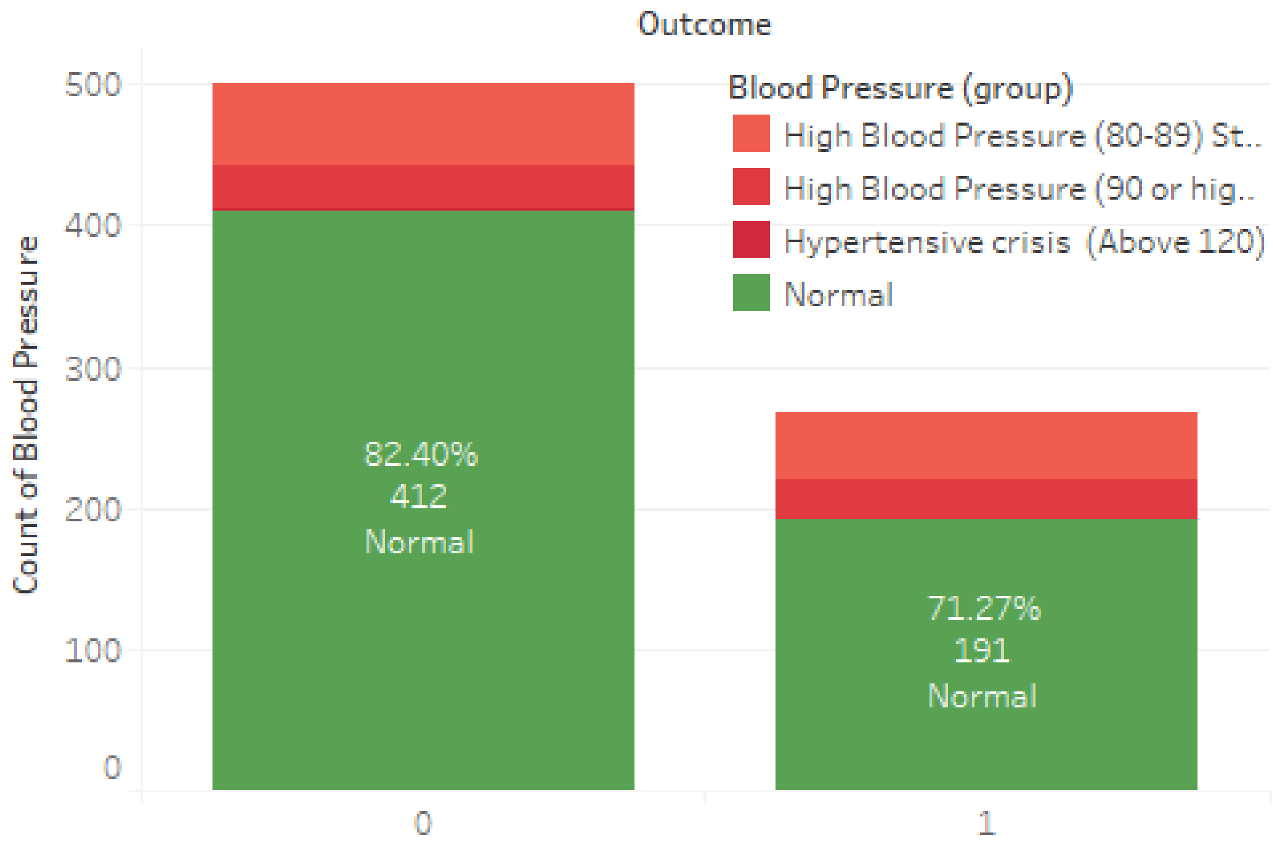
Relation between obesity and diabetes



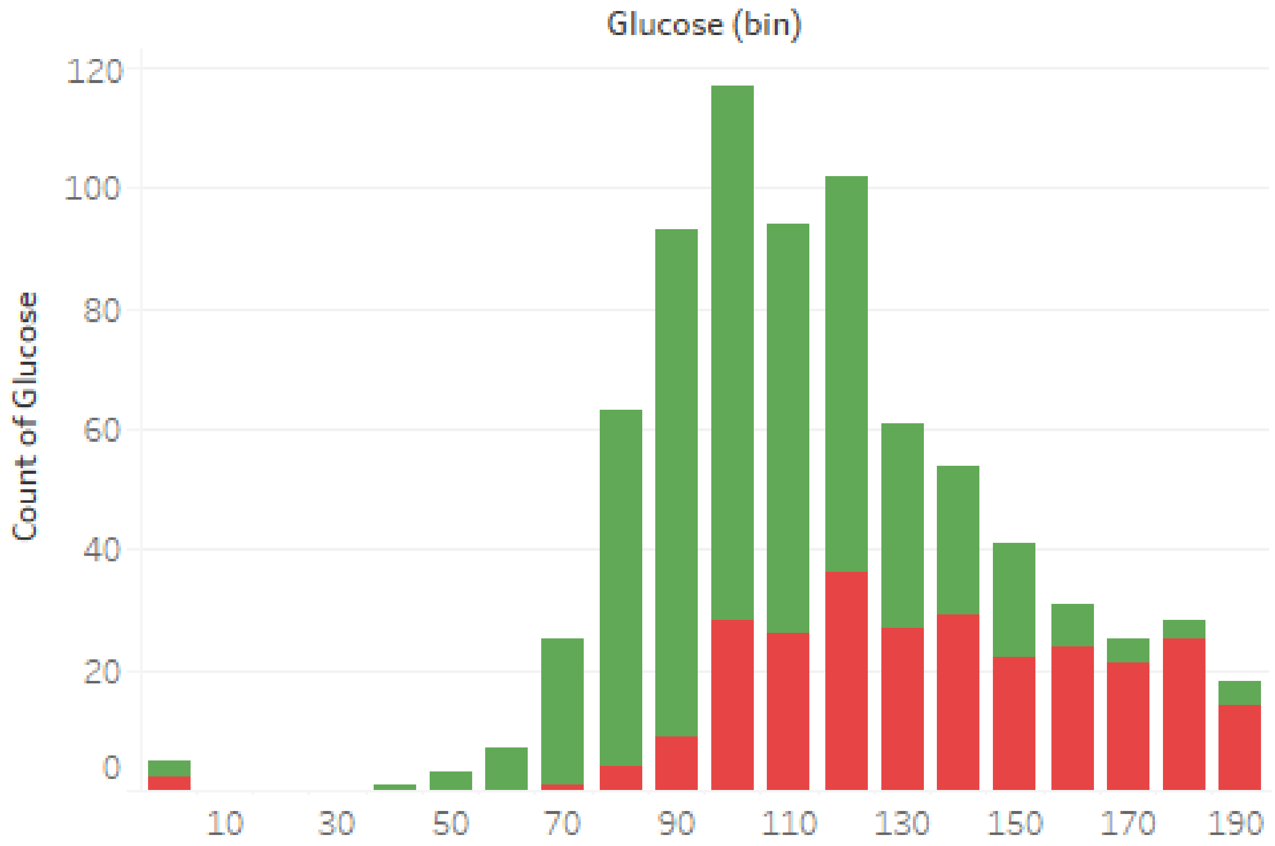
Pima indian diabetes (EDA part 2)

Pima indian diabetes (EDA part 2)

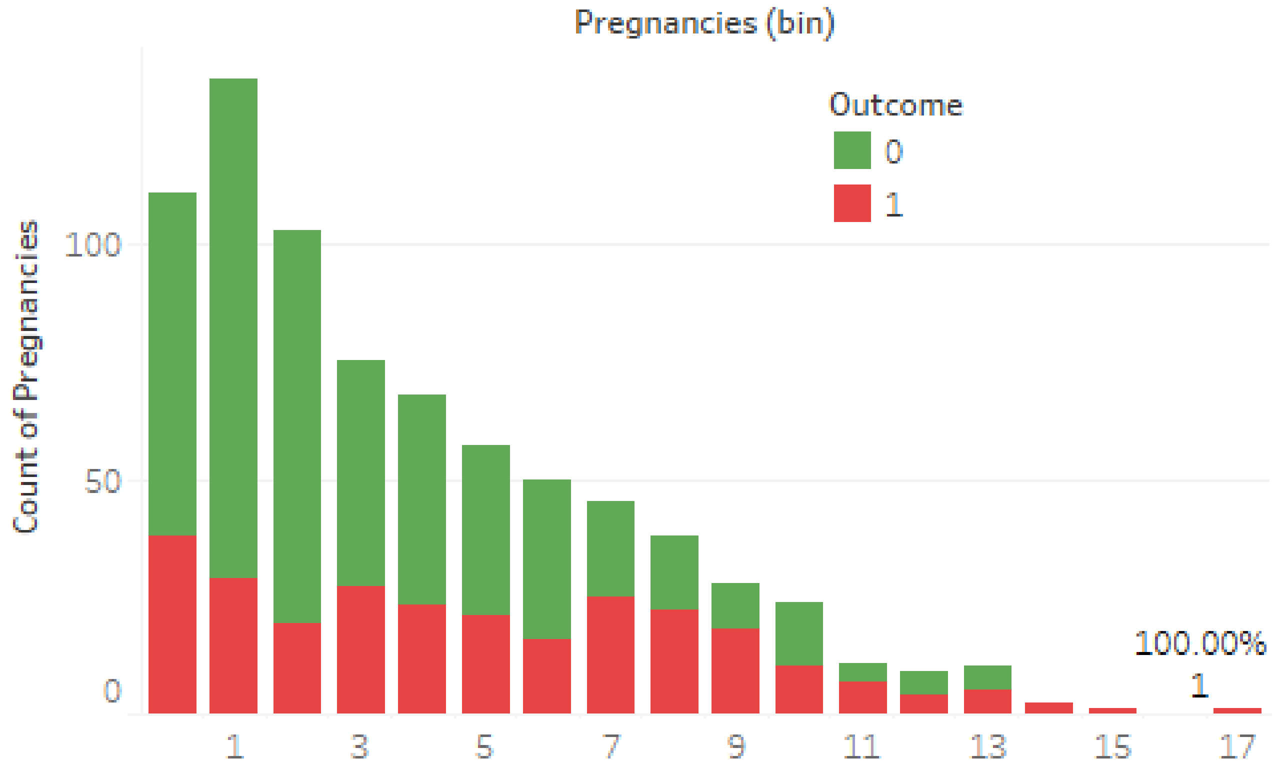
Relation between blood pressure and diabetes



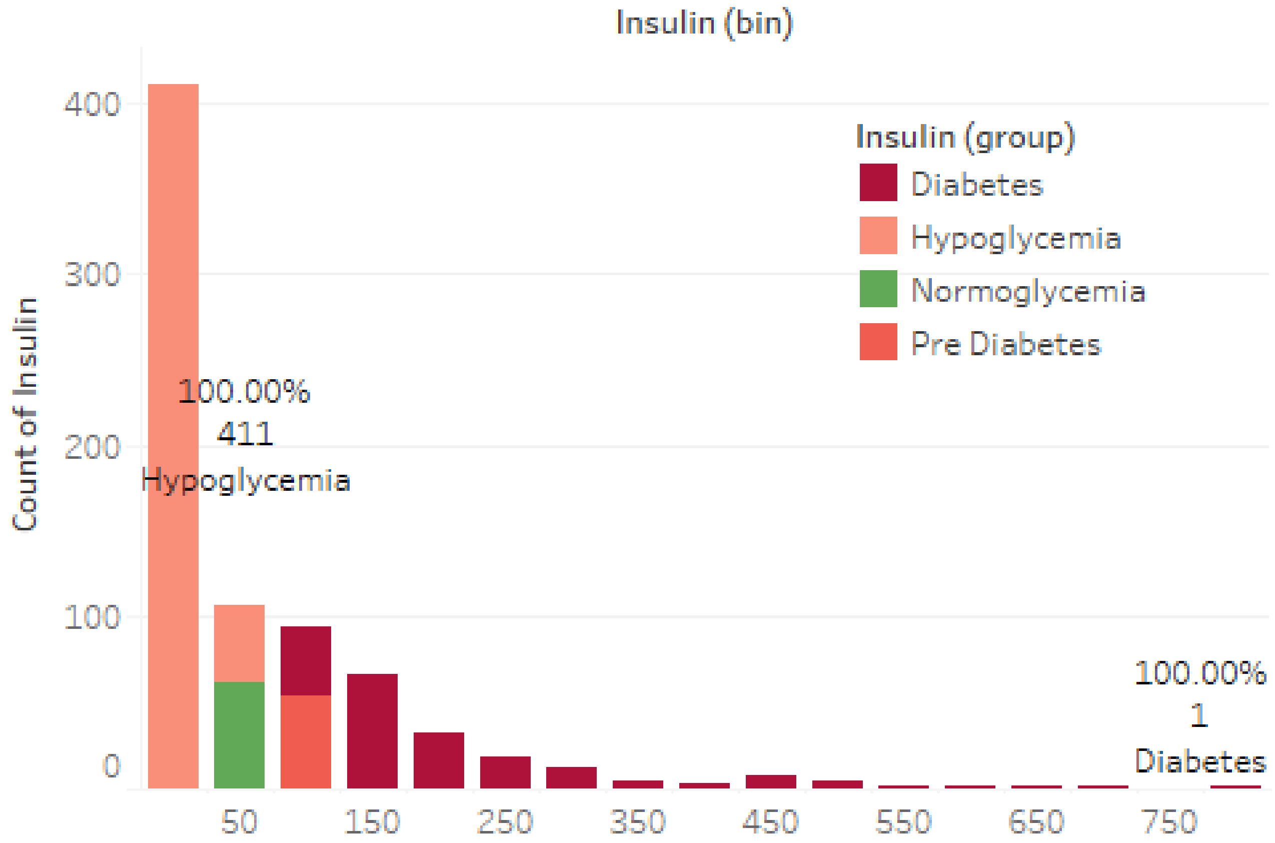
Relation between glucose and diabetes



Relation between number of pregnancies and diabetes



Relation between insulin and diabetes



Why use Machine Learning ?

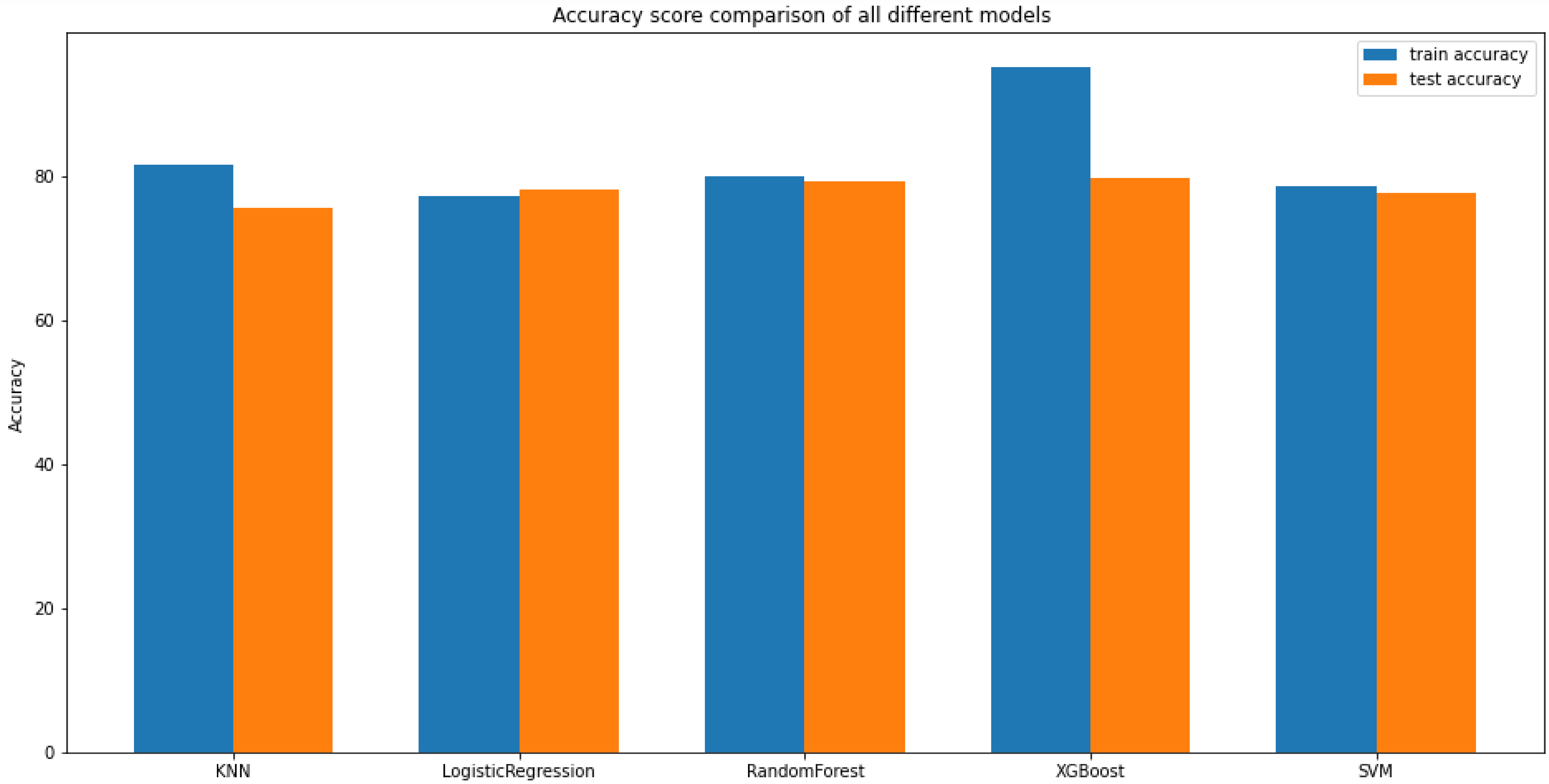
- Machine Learning will provide us the computational ability to process large data sets, use the data in different models, train the models using existing data to predict the test data and then perform evaluation metrics to see the accuracy of the model and if it is a good fit for real time data.

- It can improve the accuracy of the current system. Accuracy is the most important thing in healthcare. Being precise and quick can save many lives and using machine learning in healthcare will help us achieve it.

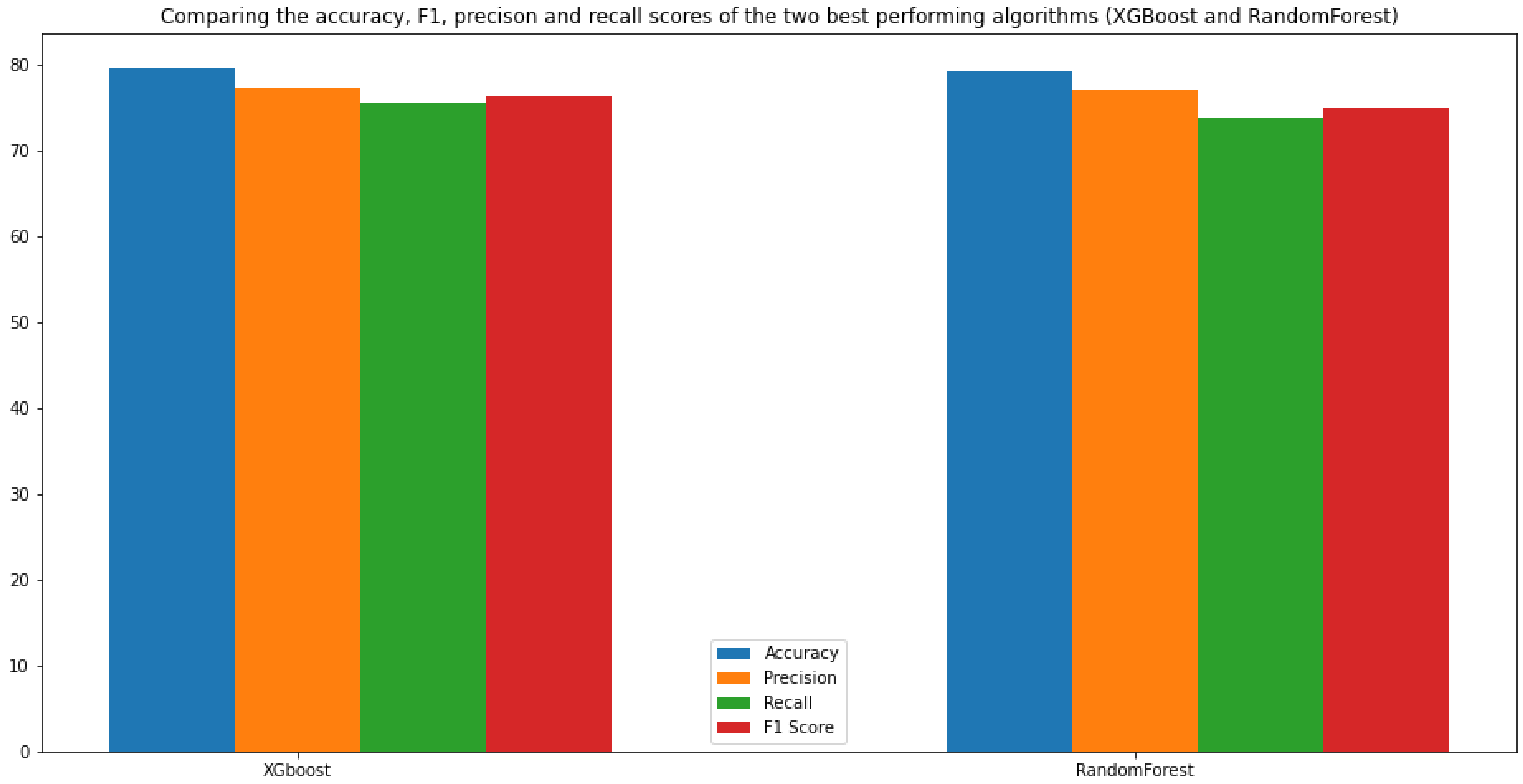
Comparing accuracies of all the algorithms

Train and Test Accuracies

Machine Learning Models	Diabetes Train Accuracy	Diabetes Test Accuracy
Logistic Regression	77.08%	78.12%
Random Forest	79.86%	79.17%
XGBoost	95.14%	79.69%
Support Vector Machine (SVM)	78.47%	77.60%
K-Nearest Neighbors (KNN)	81.60%	75.52%


Accuracy score comparison of all different models
train accuracy
test accuracy
Accuracy
0
20
40
60
80
KNN
LogisticRegression
RandomForest
XGBoost
SVM

Comparing the accuracy, F1, precision and recall scores of the two best performing algorithms (XGBoost and RandomForest)


Comparing the accuracy, F1, precison and recall scores of the two best performing algorithms (XGBoost and RandomForest)
80
70
60
50
40
30
20
10
0
XGboost
RandomForest
🟦 Accuracy
🟧 Precision
🟩 Recall
🟥 F1 Score

Thank you

References

- Kaggle Data source link: https://www.kaggle.com/uciml/pima-indians-diabetes-database
- https://en.wikipedia.org/wiki/Insulin
- https://www.diabetes.ca/media-room/press-releases/one-in-three-canadians-is-living-with-diabetes-or-prediabetes,-yet-knowledge-of-risk-and-complicatio
- https://www.freepik.com/
- https://www.canva.com/
- https://www.who.int/features/factfiles/diabetes/en/
- https://www.healthline.com/health/diabetes#:~:text=Diabetes%20mellitus%2C%20commonly%20known%20as,the%20insulin%20it%20does%20make.