DSC 630 – Predictive Analytics

Final Project

Krystina Moses

Bellevue University

Final Project

Executive Summary

Diabetes is a chronic disease in which there are three main types, Type 1, Type 2, and gestational diabetes. There are more than 122 million Americans living with diabetes today. Diabetes is the seventh leading cause of death with ninety to ninety-five percent of cases being Type 2. The Center for Disease Control and Prevention states that "Type 1 diabetes is thought to be caused by an immune reaction". Type 1 factors include family history and age. Type 2 diabetes risk factors are having prediabetes, overweight, age, immediate family history, history of gestational diabetes, and also some cultures have a higher risk. Prediabetes and Gestational diabetes have some of the same risks as Type 2, such as overweight and immediate family history. Gestational diabetes is having diabetes while pregnant. With all the listed factors, I asked myself the question of what is the likelihood that someone with all or some of the factors would develop diabetes? The data used for this project was from kaggle.com and has data such as glucose, BMI, age, number of pregnancies, and blood pressure.

The data was analyzed and models were created to determine the correlations between the factors. The results of the analysis determined that Glucose is the highest determining factor. There are preventable measures that can be taken to stay healthy. These measures include, lifestyle changes, eating healthier, and physical activity.

In my review of the data, it was determined that Glucose level had the most importance when it came to determining the likelihood of diabetes. This was followed by BMI and age. Prevention of Type 2 diabetes is possible. As the risk additionally increases with age, men are more likely to become undiagnosed as well. There are many lifestyle

changes programs that assist with prevention of Type 2 diabetes and the CDC has many great resources for this as well.

Technical Report

Introduction and Background

This project looks at diabetes and determines the likelihood of diabetes development based on several factors. The objective is to build a model in which can predict if a person has diabetes and to interpret the results to find factors that influence the outcome. There are a few different types of diabetes known and each one treated in its own way. While some may not be preventable, this disease impacts the lives of many each year with diagnosis, complications, and death. There are many that go undiagnosed as well. There is information out there that Type 2 diabetes can be prevented with lifestyle changes and the lifestyle changes are also important when diagnosed with Gestational diabetes.

Based on a recent statistics report, 10.5% of the US Population is diagnosed with diabetes and 21.4% is undiagnosed. More specific numbers include, 34.2 million people that have diabetes and 1 in 5 that are unaware of it. 88 million adults also have prediabetes. There are some reports that link the percentage of adults with obesity to having diabetes. Looking at the years from 1994 - 2015, there is a vast increase in the number of adults with both obesity and diabetes across the United States. I looked at additional comparisons with other factors that are connected and how it is all impacted.

Diabetes is a disease in which occurs when your blood sugar is too high. Nearly 1.6 million Americans have type 1 diabetes and it occurs at every age. Type 2 diabetes is the most common and a healthy diet as well as fitness are key parts to managing it. Additionally, gestational diabetes is treatable and happens to millions of women. Prediabetes does not have any clear symptoms but it can develop into type 2 so it is important to be aware as soon as possible.

The data has factors based on glucose levels, blood pressure, number of pregnancies, and BMI to determine if they have diabetes. The dataset focuses on women from the ages of twenty-one to eighty-one.

Methods

Data cleaning, exploratory data analysis, and models were created and analyzed using Python and R. The data source was from Kaggle, https://www.kaggle.com/uciml/pima-indians-diabetes-database. The target variable was the outcome of diabetes or no diabetes. The other variables included in the dataset were the number of times being pregnant, glucose, blood pressure, skin thickness, insulin, body mass index, diabetes pedigree function, and age in years.

During the data cleaning phase, it was determined there were no missing values and Exploratory data analysis showed that glucose is the most correlated with the outcome followed by BMI, see Figure 1.

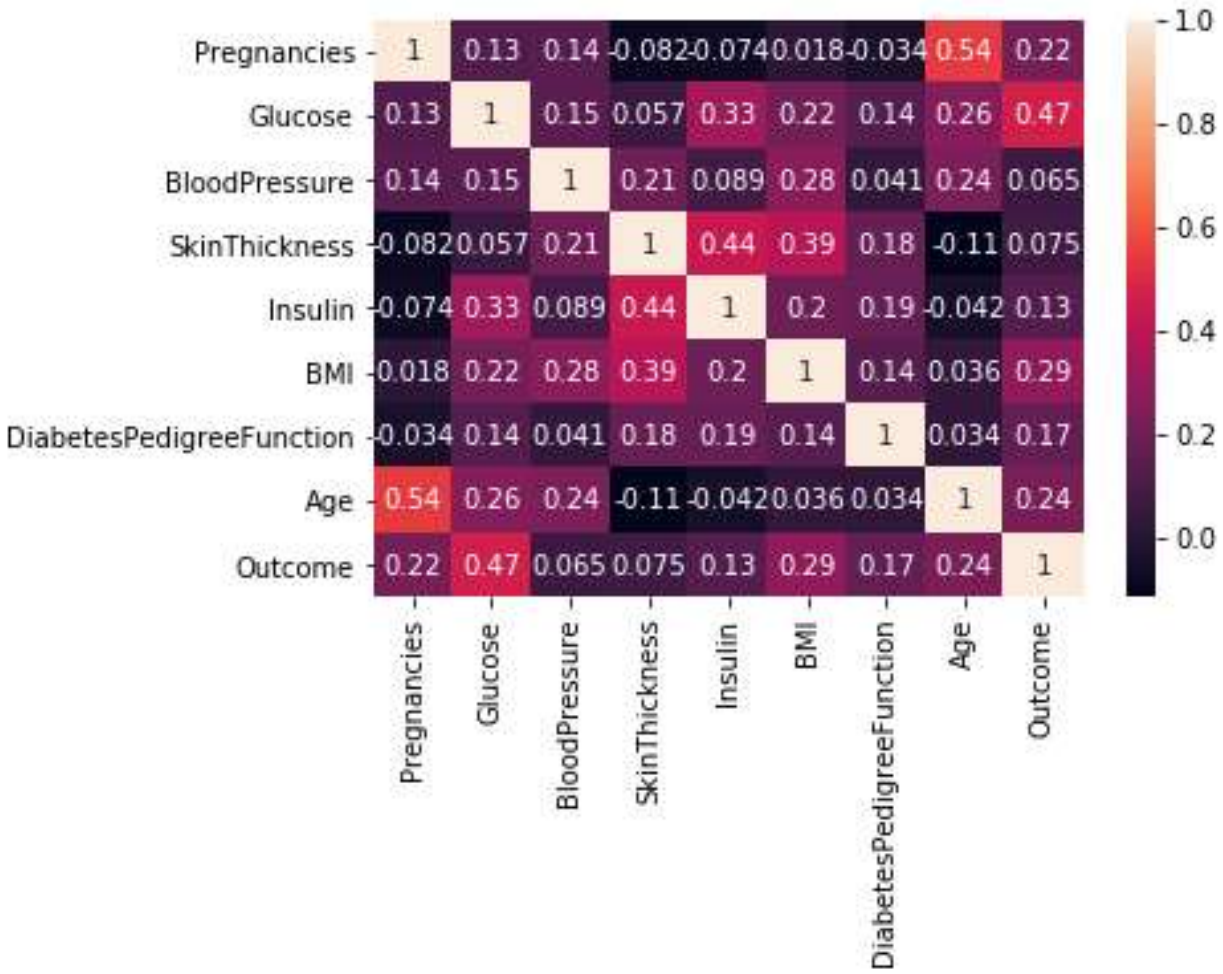
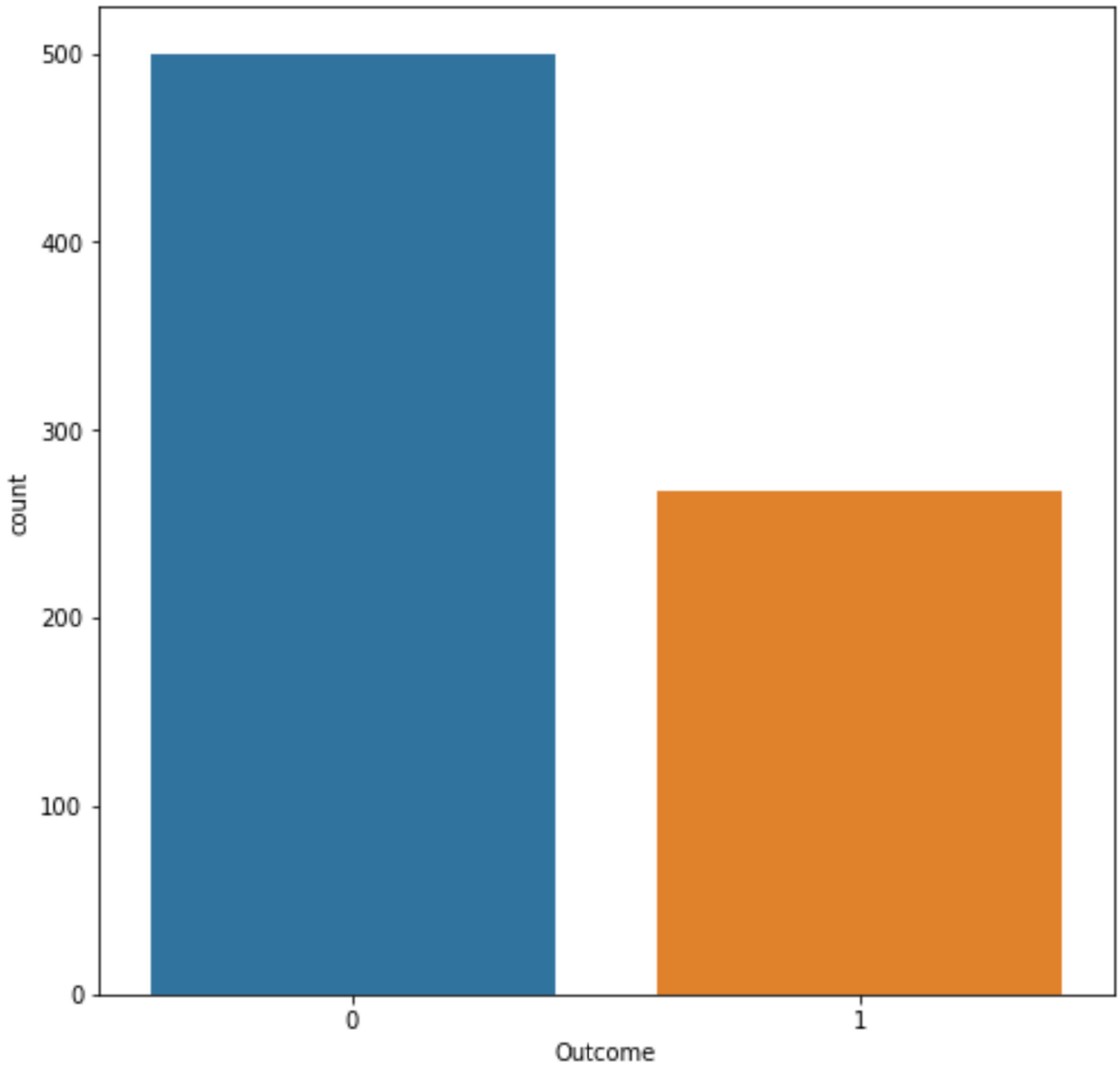
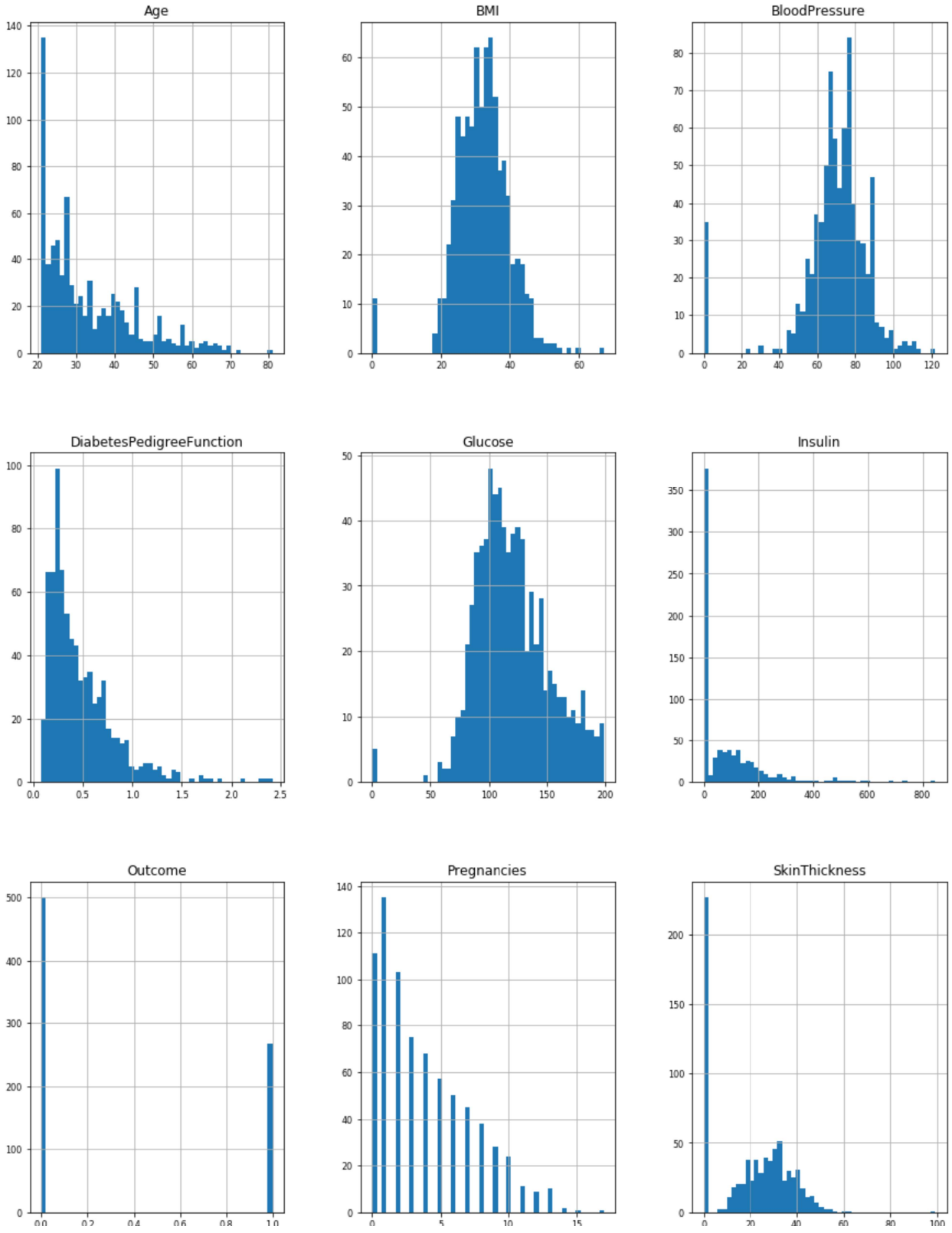


Figure 1: Correlation

This shows that if someone has a higher glucose level, their chance of diabetes goes up, same with BMI. The analysis of the number of pregnancies versus the outcome varies. The number of pregnancies correlates fifth with the outcome. The less pregnancies there are, the outcome of having diabetes is higher based on the analysis with pregnancies and outcome. With the age factor, it is third for correlation. Between the ages of twenty-one and eighty-one, it shows more diabetes outcomes in the younger age groups with twenty-five having the most.



In figure 2, we can see the outcome count of the data. This shows non diabetic represented as 0, with a total of 500 and diabetic represented as 1, with a total of 268. Figure 3 shows all variable counts.



In the following figures 4-9, it is shown with the variables and their outcomes.

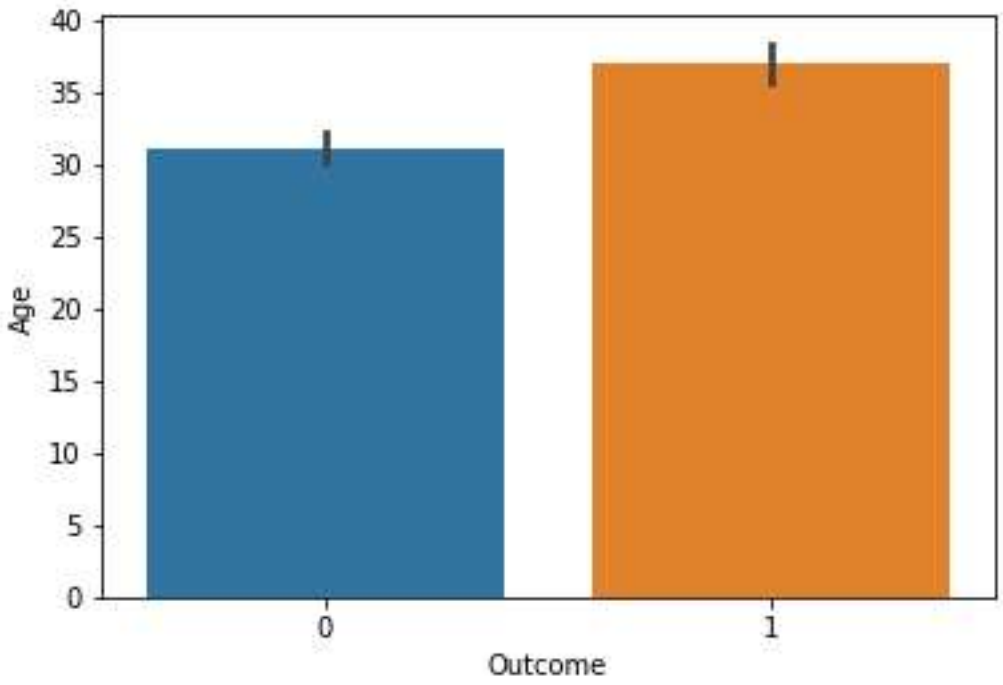


Figure 4: Age/Outcome

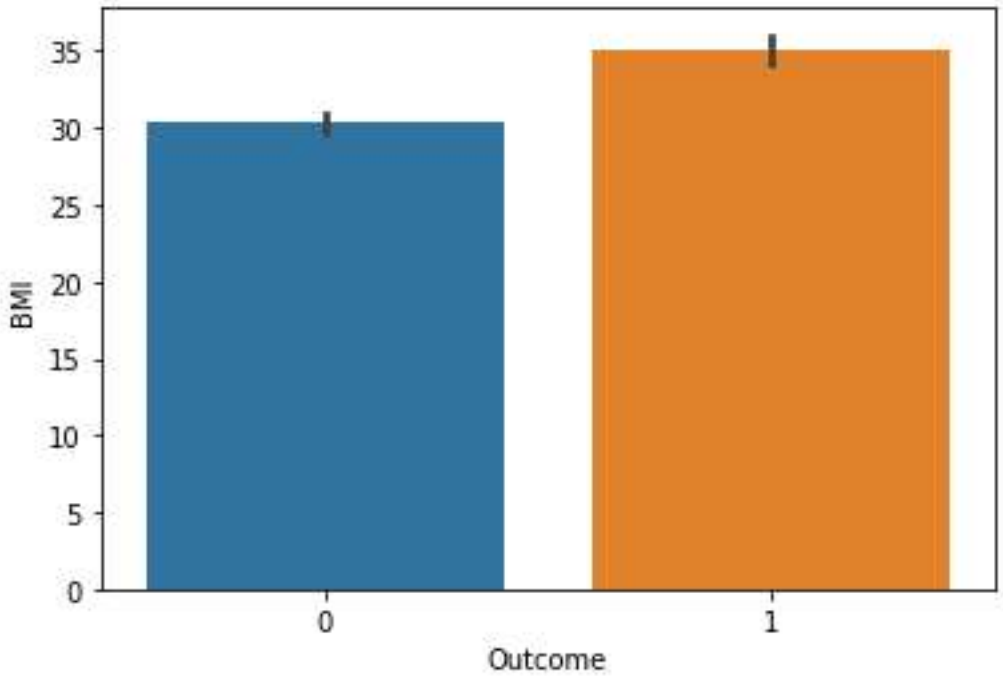


Figure 5:BMI/Outcome

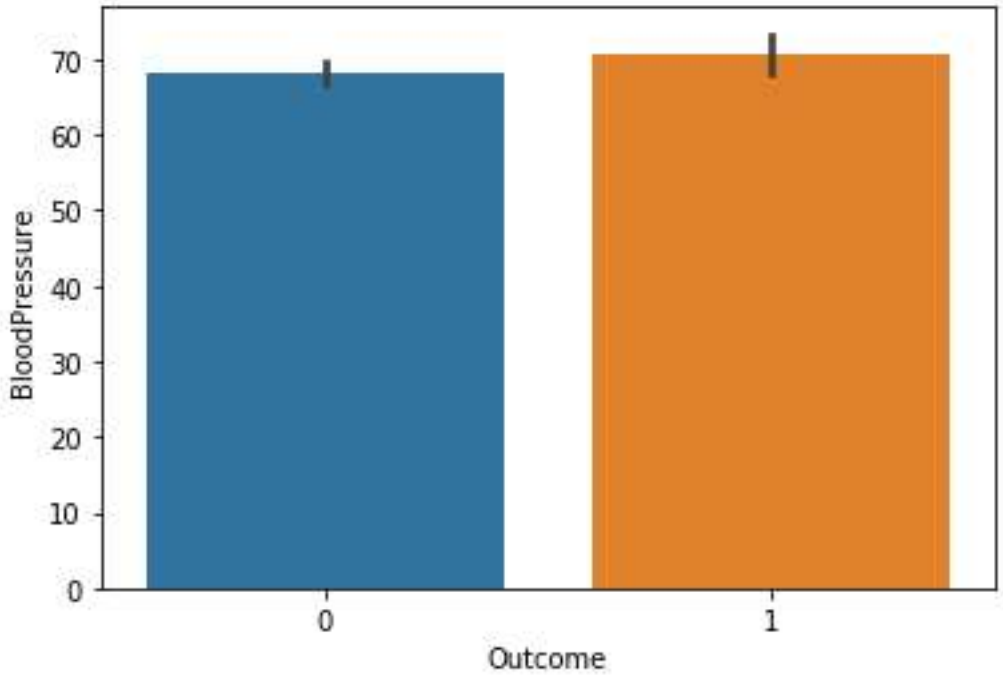


Figure 6:Blood Pressure/Outcome

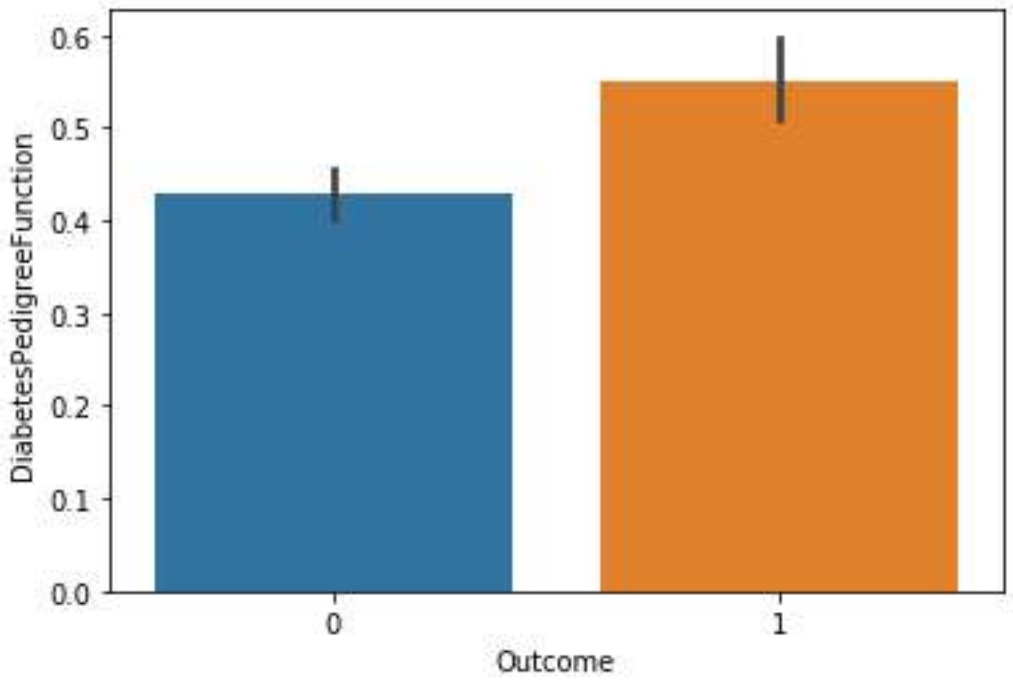


Figure 7:Diabetes Pedigree Function/Outcome

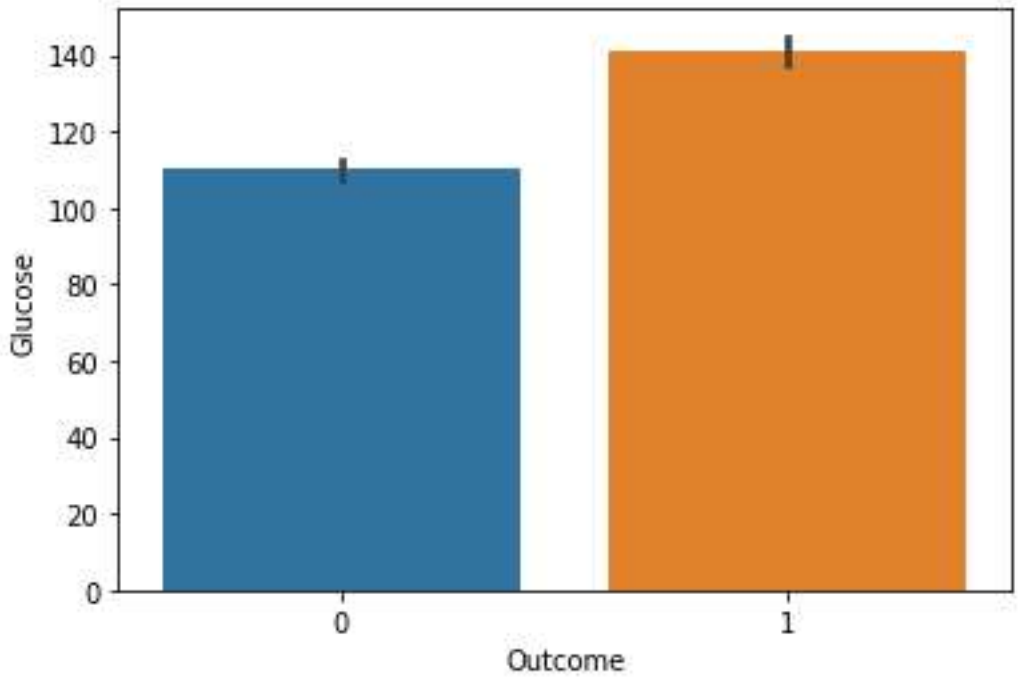


Figure 8:Glucose/Outcome

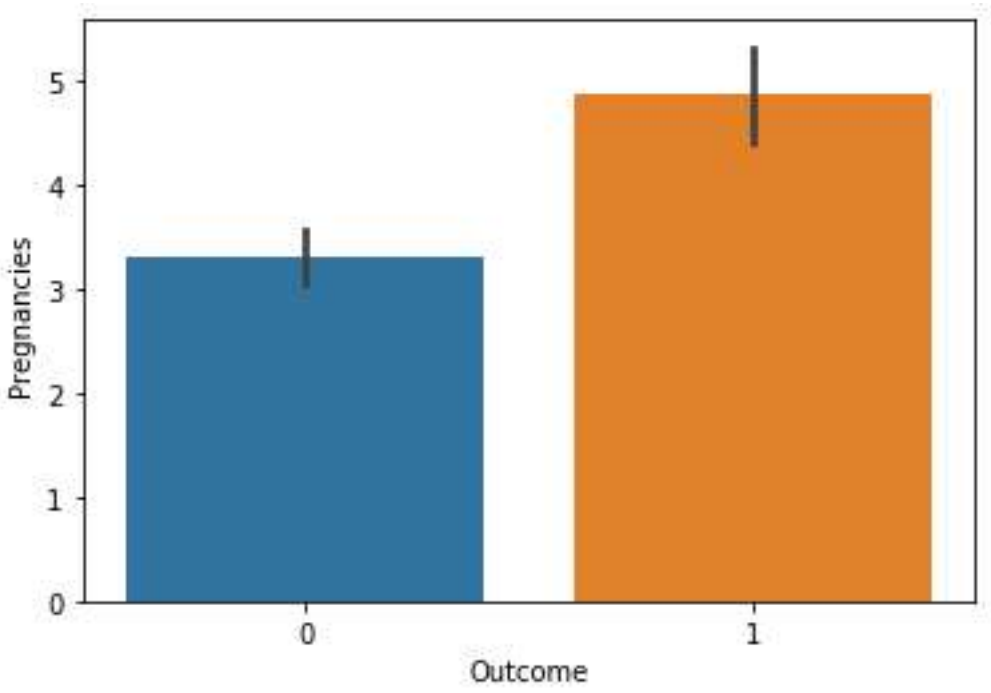
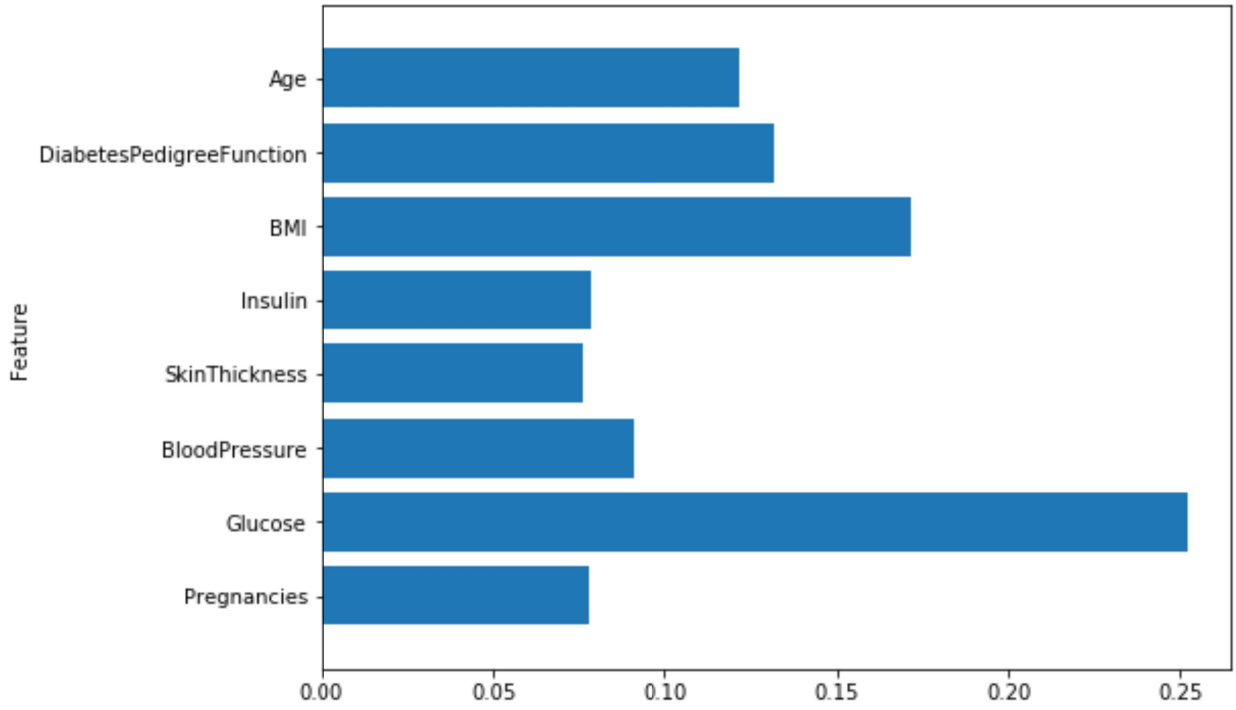


Figure 9:Pregnancies/Outcome

I worked with different models to determine the best one. I started with linear regression with an 80/20 percent split with the training and testing data. This came out with an accuracy of 29%. I additionally used the 80/20 percent training and testing data for logistic regression, support vector classifier (SVC), and random forest. In the following table the information gathered from the models is shown.

	Logistic Regression	SVC	Random Forest
F1 Score	0.38	0.39	0.42
Precision Score	0.32	0.33	0.43
Recall Score	0.47	0.50	0.47
Accuracy	0.62	0.65	0.59

Results



With the models that I ran, it was determined that the SVC model came out with the best accuracy. We could also go based on the other scores as well such as recall which is defined by towarddatascience.com as "the precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives." This would still give us the result of the SVC model.

The results determined that Glucose has the highest importance when determining if a person has diabetes or is at risk of diabetes.

Discussion/Conclusion

A lot of knowledge was gained around this subject during the course of this project and much of the prevention for diabetes consists of lifestyle changes. High glucose levels, older age, and a higher body mass index are big contributors to diabetes. Additionally, as a woman, adding in pregnancies and the amount of them are a factor as well. There is a more of a likelihood for diabetes during and after. Type 2 diabetes is the most common and can change your life. The importance of this project for me was to determine the most

outstanding factor as well as those leading up to it. Glucose, as I suspected, was the highest factor. Based on additional observation, blood pressure, and diabetes pedigree function are not low on the scale either.

Acknowledgments

I would like to thank the American Diabetes Association and the Center for Disease Control and Prevention for the information provided. I would not know as much about it without their assistance. I would also like to thank Professor Kern and my classmates for the feedback provided throughout the quarter to better improve myself and my work.

References

Diabetes. (2020, February 18). Retrieved from https://www.cdc.gov/diabetes/library/socialmedia/infographics/diabetes.html

Diabetes Overview. (n.d.). Retrieved from https://www.diabetes.org/diabetes

Diabetes Prediction - dataset by informatics-Edu. (2019, August 27). Retrieved from https://data.world/informatics-edu/diabetes-prediction/workspace/file?filename=Diabetes+Registry.CSV+-+Introduction+to+Biomedical+Data+Science.csv

Gestational diabetes. (2020, February 27). Retrieved from https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/symptoms-causes/syc-20355339

National Diabetes Statistics Report. (2020, February 14). Retrieved from https://www.cdc.gov/diabetes/data/statistics/statistics-report.html

U.S. Diabetes Surveillance System. (n.d.). Retrieved from https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html#