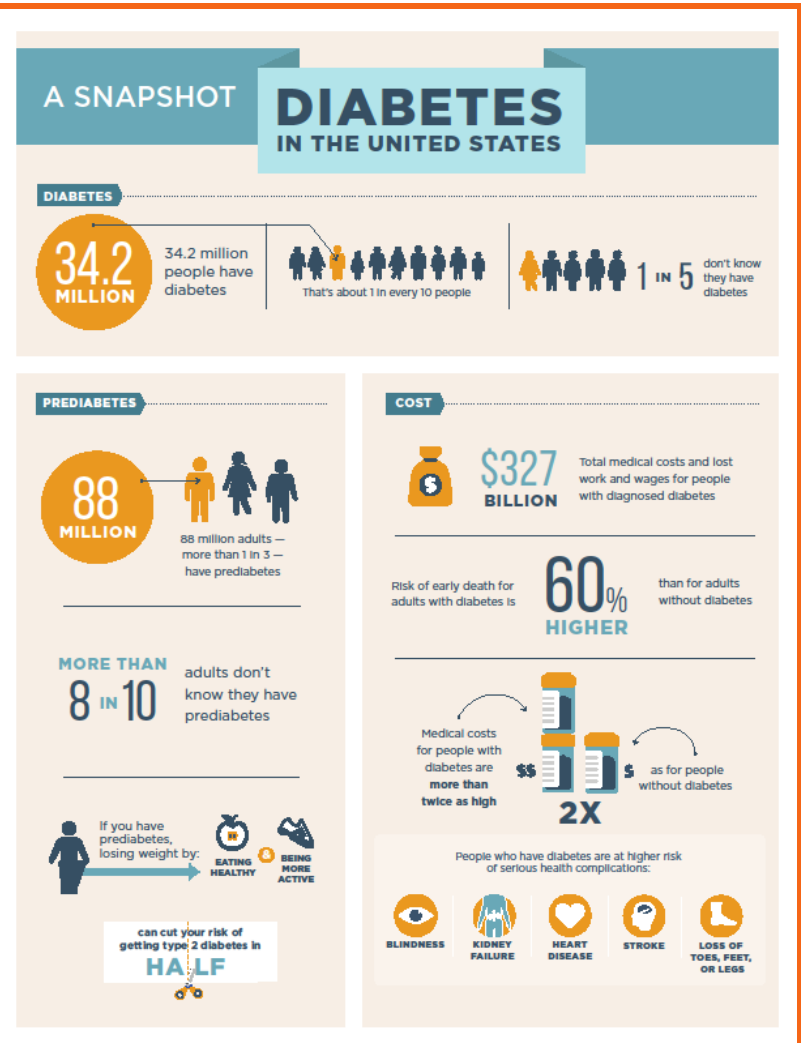
PREDICTION OF DIABETES

DSC 630 – Final
Krystina Moses

INTRODUCTION/BACKGROUND

- This project looks at diabetes and determines the likelihood of development as well as how much the amount of people with the disease has changed over the years.
- There are a few different types of diabetes known and each one treated in it's 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.



DATA

- The dataset covers many factors including glucose, BMI, age, number of pregnancies, and blood pressure . It is also used to determine if a person will have diabetes.
- https://www.kaggle.com/uciml/pima-indians-diabetes-database
- A brief look at the data:

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

DATASET DESCRIPTION

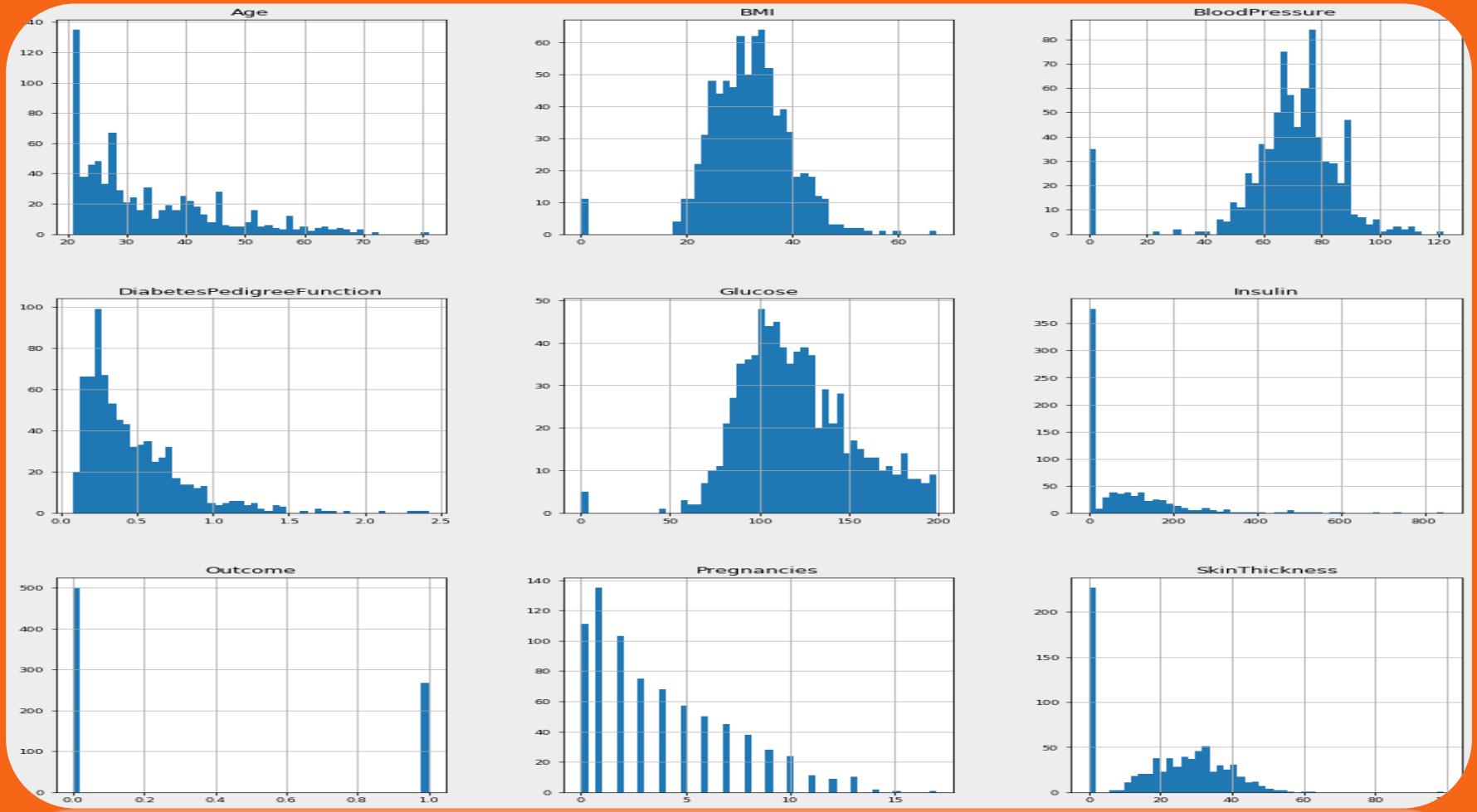
- **Pregnancies**: # of times pregnant
- **Glucose**: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- **Blood Pressure**: Diastolic blood pressure (mm Hg)
- **Skin Thickness**: Triceps skin fold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)
- **BMI**: Body mass index (weight in kg/(height in m)^2)
- **Diabetes Pedigree Function**: a function that represents how likely they are to get the disease by extrapolating from their family history
- **Age**: Age in years
- **Outcome**: Class variable (0: non-diabetic, 1: diabetic)

DATA PREPARATION / EXPLORATORY DATA ANALYSIS

- Verifying the data had no missing or NaN values
- Renaming columns to ensure they were all the same format and for easy reading
- Verifying the type of information on the dataset

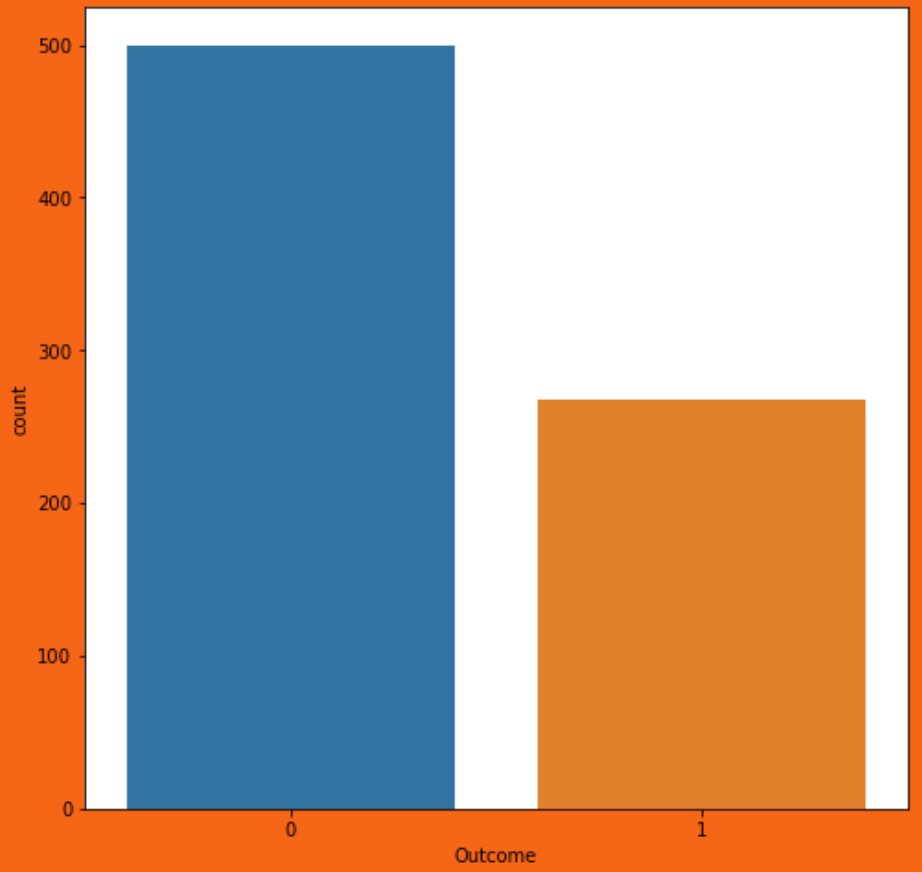
- Data visualizations to better understand the data

GRAPHS FOR EACH VARIABLE

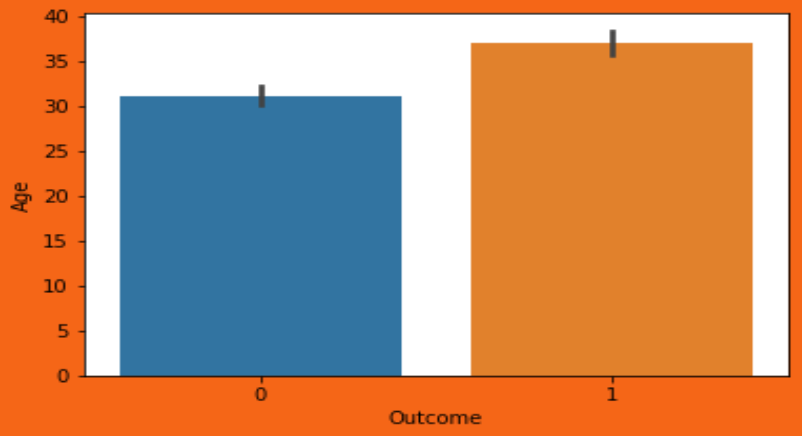


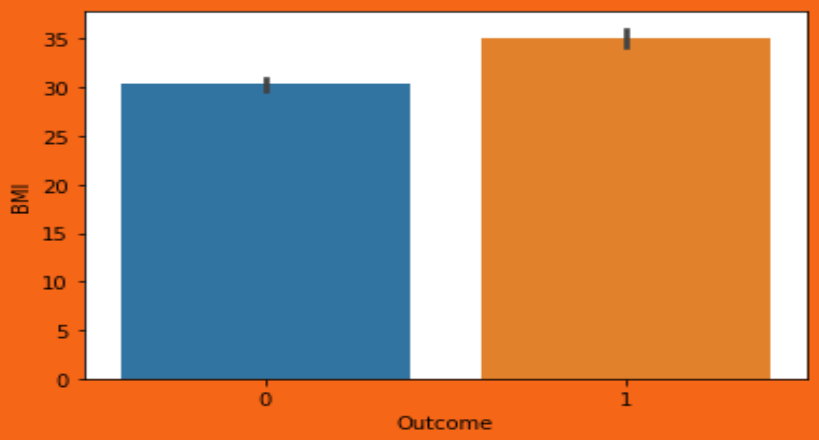
OUTCOME – TARGET VARIABLE

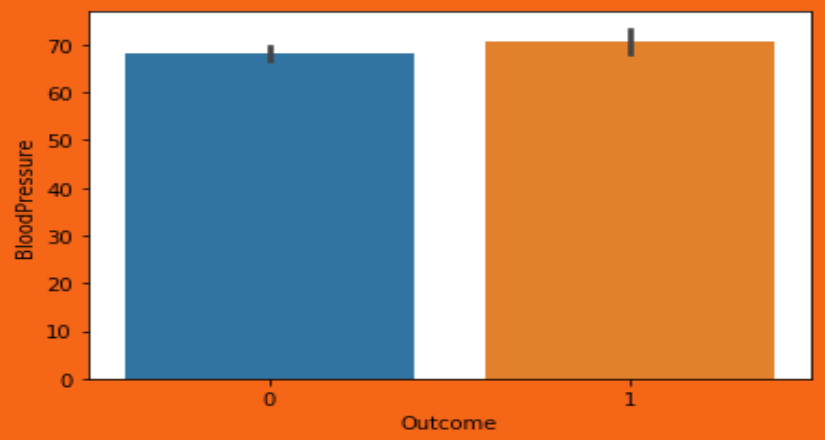
- Non diabetic is represented as zero and has 500 values
- Diabetic is represented as one and has 268 values

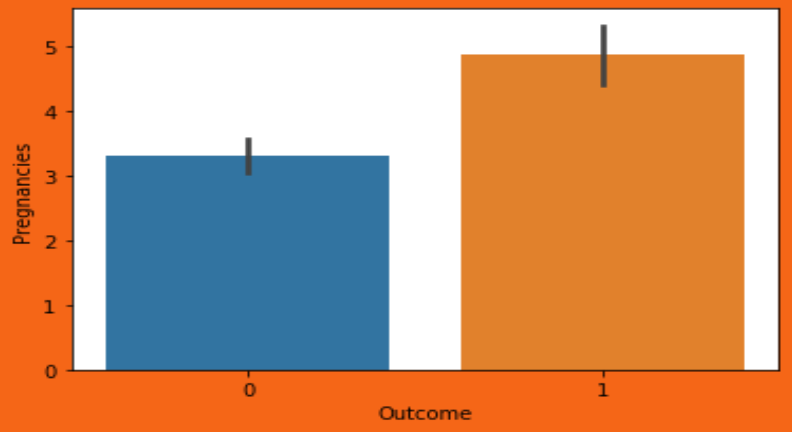


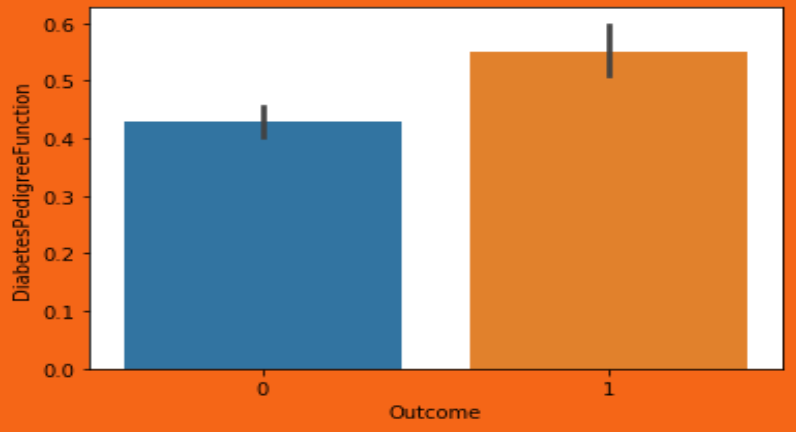
OTHER VARIABLES IN RELATION TO OUTCOME

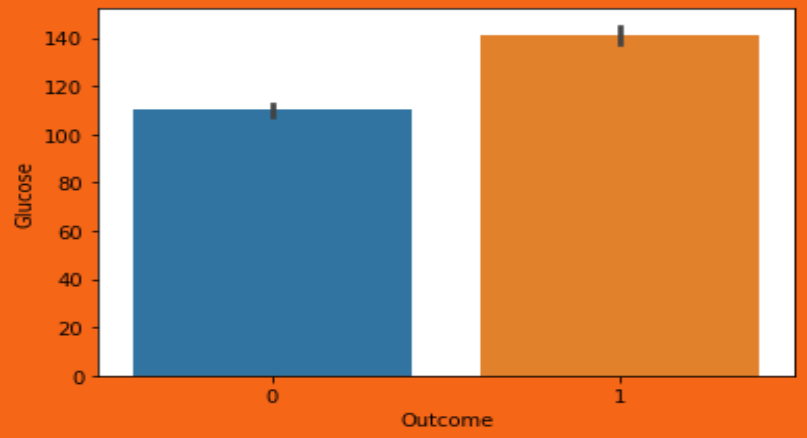




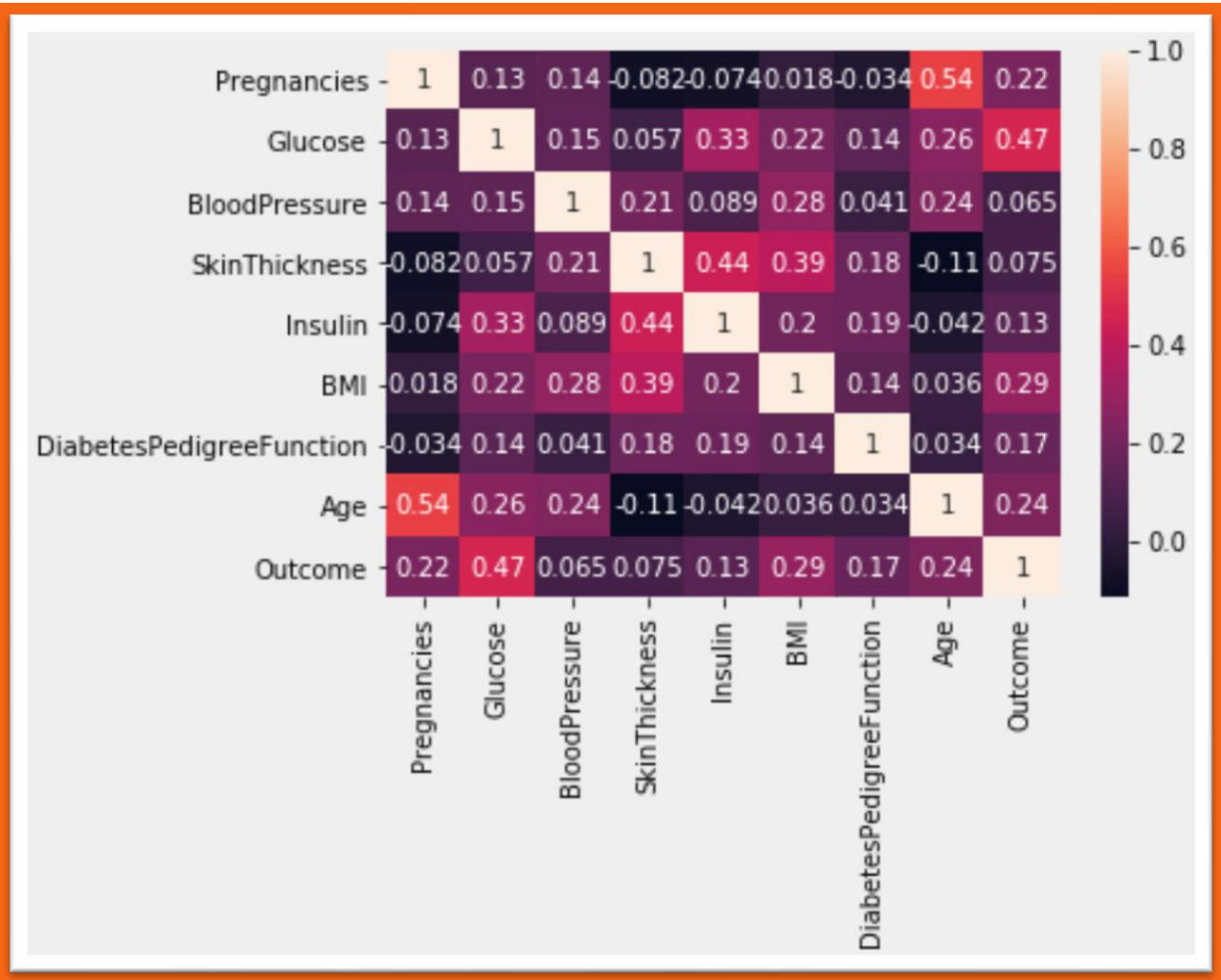








UNDERSTANDING THE DATA



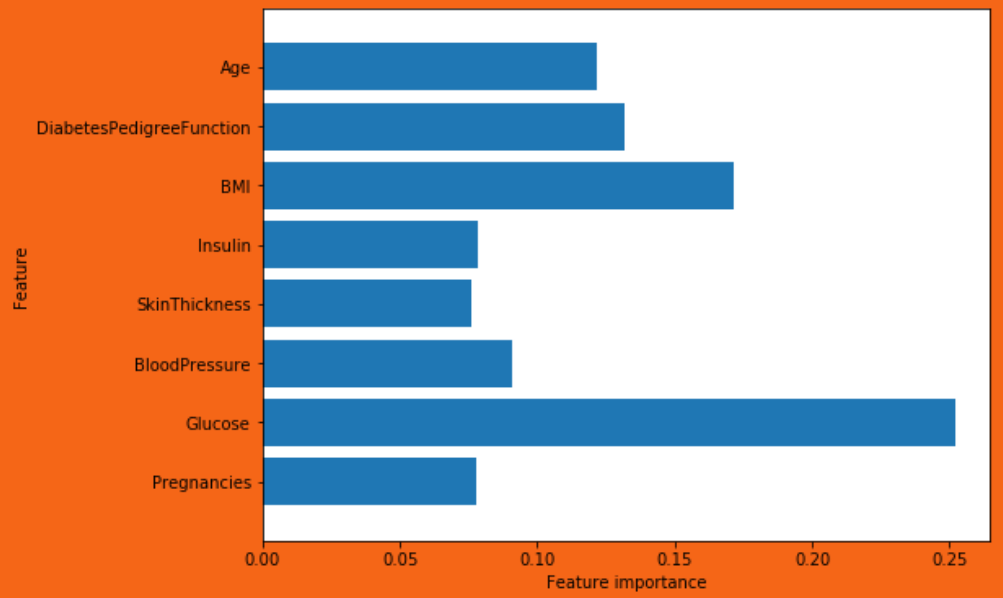
A correlation matrix was created to see the relationship between the variables. This determined that Glucose had the closet relationship to a person who had diabetes with BMI as the next closest.

MODELS

	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

- Use models to determine the likelihood of diabetes and the most common factor
- Support Vector Classifier came out with the best accuracy.

RESULTS/CONCLUSION



- It was determined that Glucose has the highest importance in determining if one has diabetes.
- This is also the case for Age, BMI, and Diabetes Pedigree Function.

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#