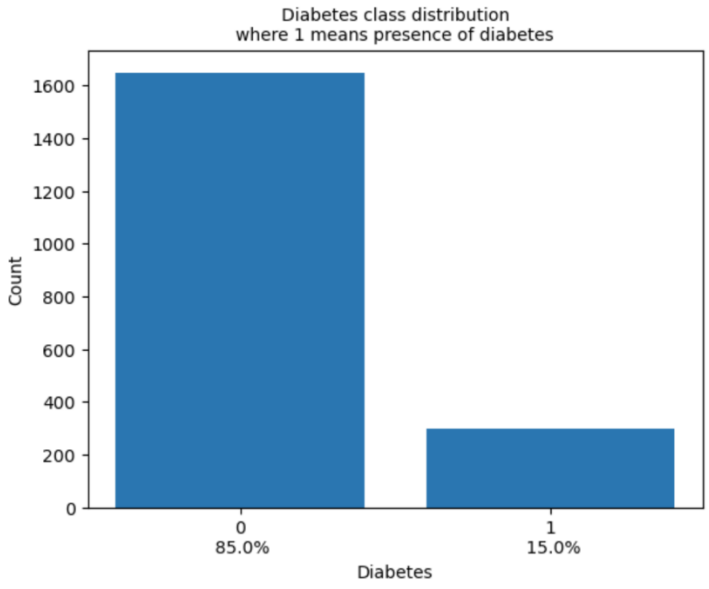
Predicting Diabetes from Patient Health Factors

Problem Statement:

We aim to look at patients' physical and health factors to predict diabetes. The goal is to create a model that can predict a person has diabetes and recommend they get tested for diabetes. This can lead to earlier intervention that may allow patients to prevent the development of diabetes.

Data Wrangling:

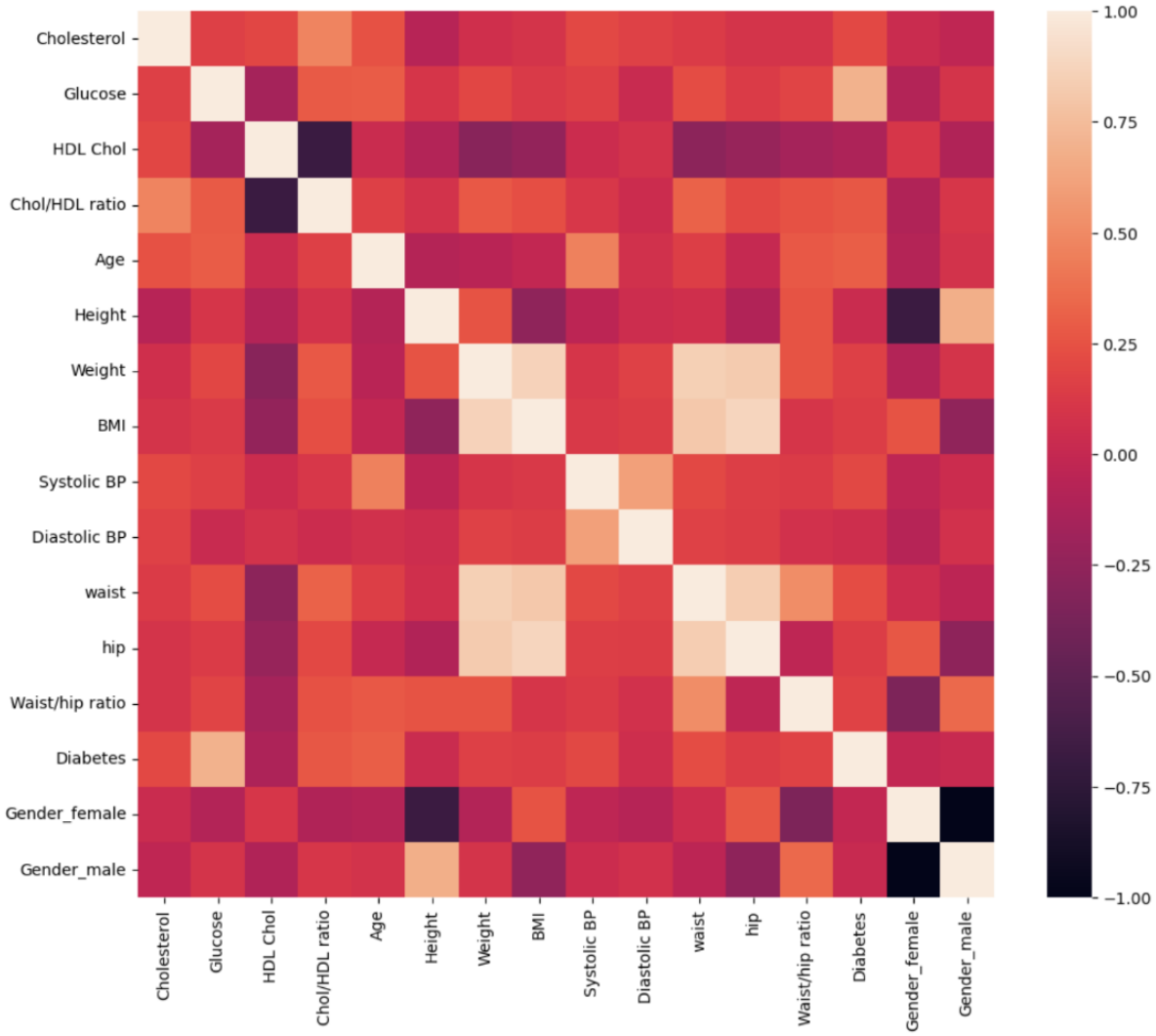
The data came from rural African-American patients. The data contained 17 columns and 390 entries. Columns included patient number, cholesterol, glucose, HDL cholesterol, chol/HDL ratio, age, gender, height, weight, BMI, systolic BP, diastolic BP, waist, hip, waist/hip ratio, and diabetes. The patient number was set to the index for the dataset. There were two unnamed columns that were removed because they were mostly null values. Removing the two unnamed columns eliminated all null values from the dataset. The target variable in this dataset is the diabetes column which is categorical and identifies patients with diabetes based on their A1c. Since the dataset only has 390 entries, the entries were multiplied to create more data. The dataset was multiplied by five, creating a dataset with 1950 entries. The class distribution of the data was 15 percent for the target variable, patients with diabetes. The remaining 85 percent were classed as not having diabetes.



Exploratory Data Analysis and Initial Findings:

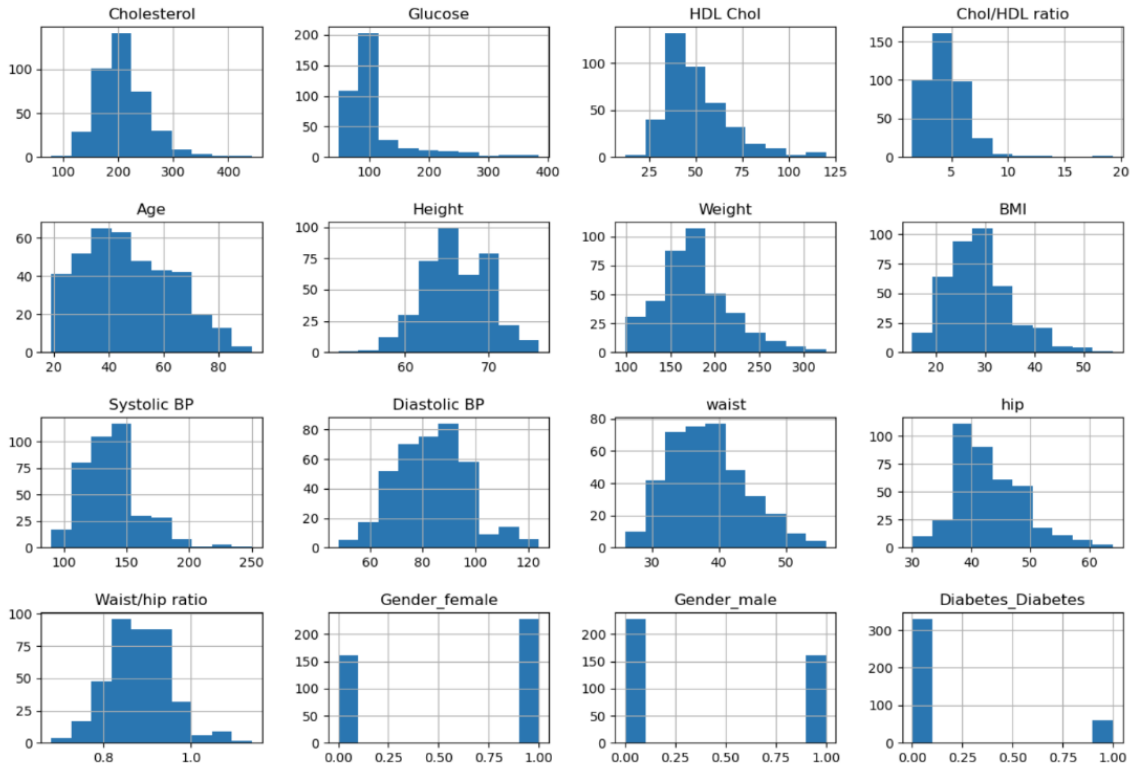
In the exploratory data analysis, I first looked at the data types for the variables. There were 13 numerical columns and two categorical variables. The numerical columns included cholesterol, glucose, HDL cholesterol, chol/HDL ratio, age, height, weight, BMI, systolic BP, diastolic BP,

waist, hip, and waist/hip ratio. The numerical columns had a data type of either integer or float. The categorical columns included gender and diabetes. Since diabetes is the target variable, this column was changed to numeric. In the diabetes column, 1 signified a patient with diabetes and 0 signified a patient without diabetes. Gender was split into gender_female and gender_male with numeric values. Below is a correlation heatmap of the dataset. Focusing on our target variable, some of the variables that appear correlated to diabetes include glucose, cholesterol/HDL ratio, age, and systolic BP.

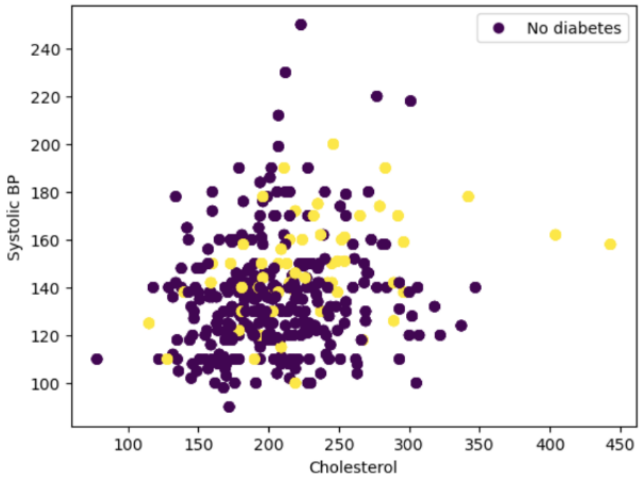


Next, I created a histogram plot for each variable to see the distribution of the data. We can see that diabetes has two values: zero and one and that there are more zeros than ones. There is a

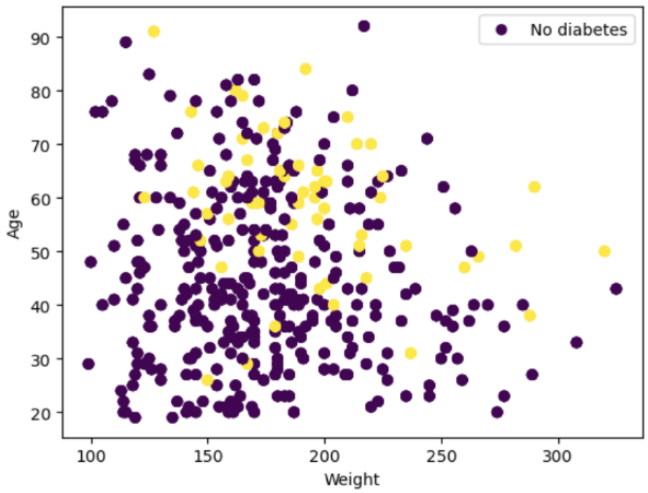
skew for glucose and systolic BP. There are more females than males in the sample, but the distribution is close to even. There are about 58% female patients and 42% male patients.



The figures below show two scatter plots. The first shows the relation between cholesterol and systolic BP. The purple dots represent patients without diabetes and the yellow dots represent patients with diabetes.

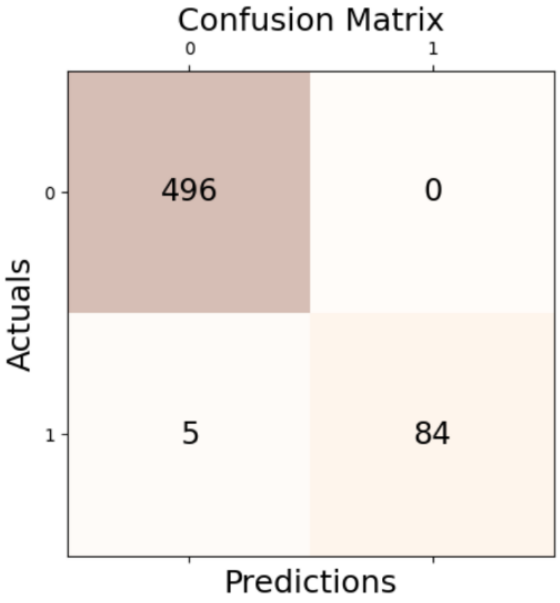


In the second figure, there relation between weight and age is shown. Again patients without diabetes are represented by the purple dots and the patients with diabetes are represented by the yellow dots.



The Model and Recommendations:

For the modeling portion, I created three different models: decision tree entropy model, logistic regression model, and a gradient boosting model. I chose to go with the logistic regression model because it had the best model statistics. The model statistics: accuracy equaled 0.99, precision equaled 1.0, recall equaled 0.94, F1 score equaled 0.97. I ran parameter tuning on the logistic regression model's solver and C feature. Using gridSearchCV I found that the solver equal to newton-cg and C equal to 10 were the best parameters.



With this model, doctors could predict patients that are likely to have diabetes and recommend them to get tested for diabetes. The model can also better help us recognize risk factors for diabetes. Knowing the risk factors can allow health professionals to better educate the public on preventing diabetes. One more recommendation for this model is to use it to create a survey for people to complete and determine if they are at risk for diabetes.

Ideas for Further Research:

One of the major problems with this dataset was the lack of data. In the future, I would recommend collecting more data to improve the model's performance. I think the model's use could continue to improve by adding more variables to the model. Finding more variables that correlate to diabetes can further help health professionals recommend preventative measures to patients. The data all comes from the same population, so diversifying the data sample will help to generalize the model's use to a wider range of people. Another thing that could be considered is the effect these variables have on diabetics. This model does not tell us whether these variables are caused by diabetes or if diabetes causes changes to these variables in a patient.