Rebecca Teng
Data Bootcamp

Predicting Diabetes using the Pima Indians Diabetes Database

Introduction to Problem and Data

For this project, I will develop a predictive model that determines whether the patient has diabetes based on health metrics such as glucose level, BMI, and age. Diabetes is a chronic condition that significantly impacts an individual's health and quality of life, so early detection is important to manage and prevent the condition from worsening.

This model is useful for healthcare professionals to identify diabetes in their patients early. Predicting the issue early allows for preventative measures to be implemented. In addition, we can identify the biggest predictors associated with diabetes for females of Pima Indian heritage, allowing professionals to monitor these healthcare metrics more closely.

Dataset Description

The data is from the UCI Machine Learning Repository and has information about female patients of Pima Indian heritage who are at least 21 years old. Given the amount of 0s in the dataset, it will require cleaning and preprocessing to handle potential missing or inconsistent values. There will be challenges to constructing an accurate classification model due to the complexity and variability of health data and the interdependence on certain health metrics. However, I believe that certain predictive variables can somewhat predict whether a patient has diabetes.

This dataset contains information about various Pima Indian patients and whether they have diabetes. There are features are as follows:

- Pregnancies: Number of times the patient has been pregnant
- Glucose: Plasma glucose concentration (mg/dL)
- BloodPressure: Diastolic blood
- pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: A function scoring the likelihood of diabetes based on family history
- Age: Age of the patient (years)
- Outcome: Target variable (0 = No diabetes, 1 = Diabetes)

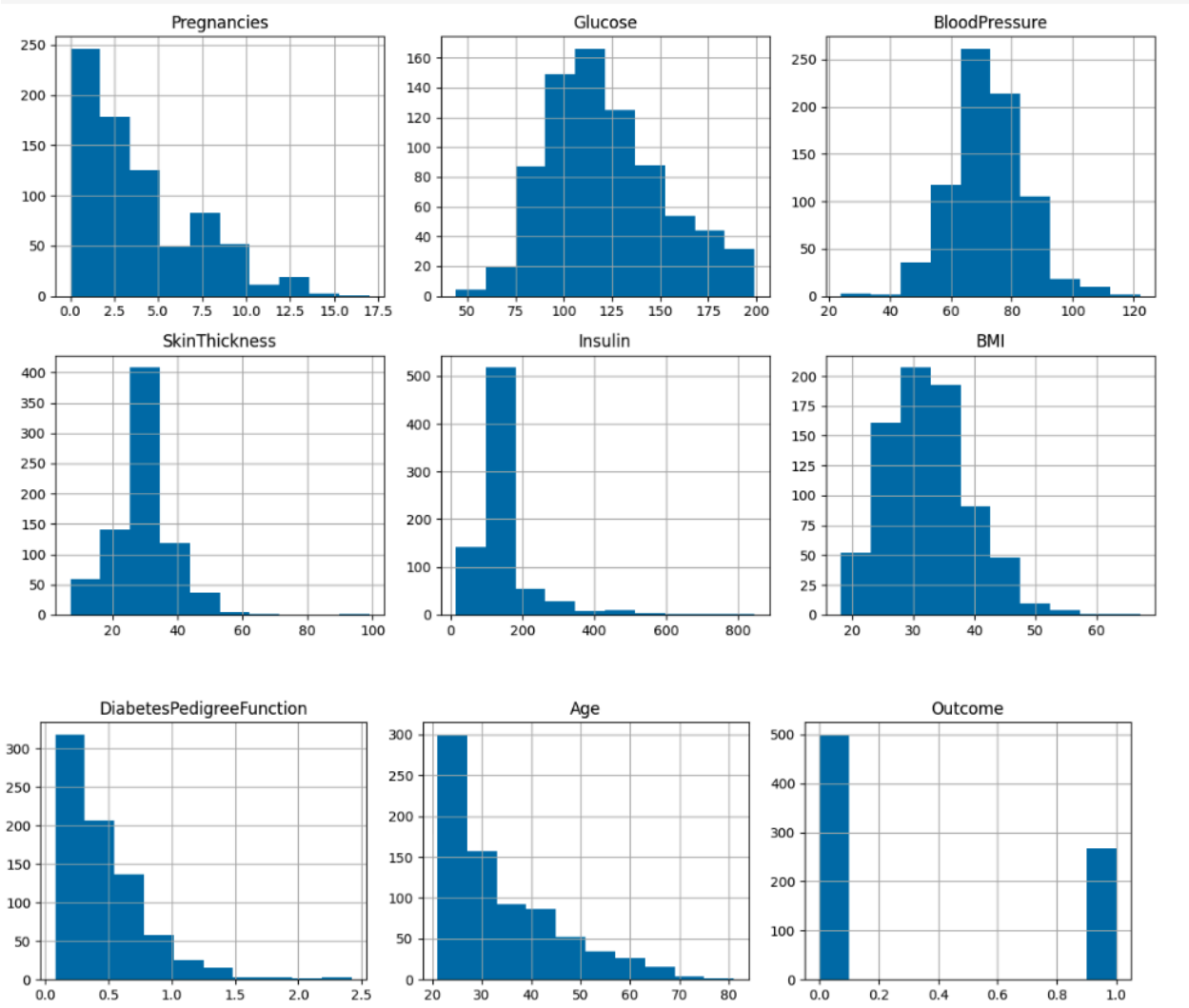
This dataset contains 768 rows of patient data and 9 columns representing various health-related attributes, which will be used to predict the likelihood of diabetes.

Descriptive Statistics

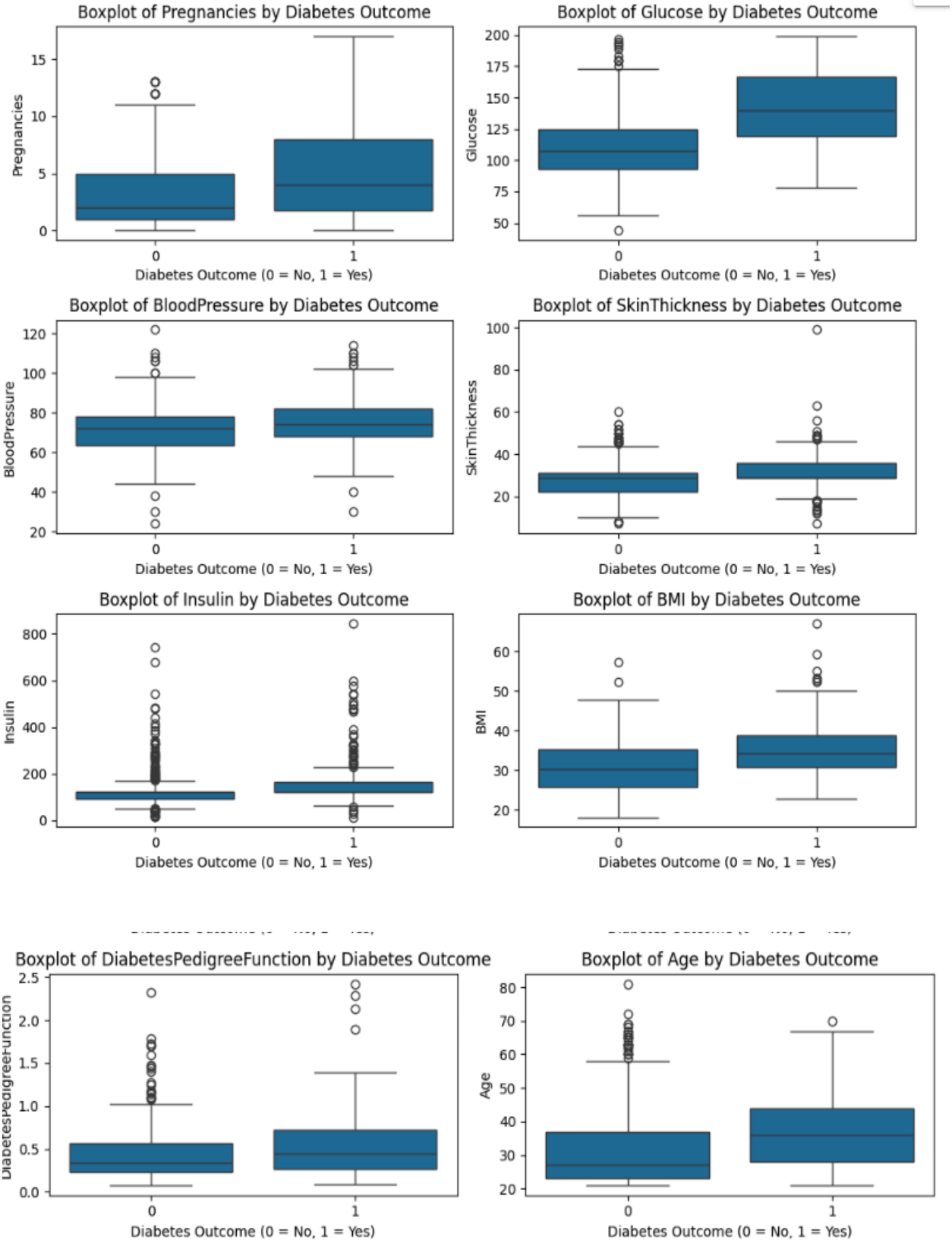
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479
std	3.369578	31.972618	19.355807	15.952218	115.244002
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000
75%	6.000000	140.250000	80.000000	32.000000	127.250000
max	17.000000	199.000000	122.000000	99.000000	846.000000

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	31.992578	0.471876	33.240885	0.348958
std	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.078000	21.000000	0.000000
25%	27.300000	0.243750	24.000000	0.000000
50%	32.000000	0.372500	29.000000	0.000000
75%	36.600000	0.626250	41.000000	1.000000
max	67.100000	2.420000	81.000000	1.000000

Exploratory Data Analysis

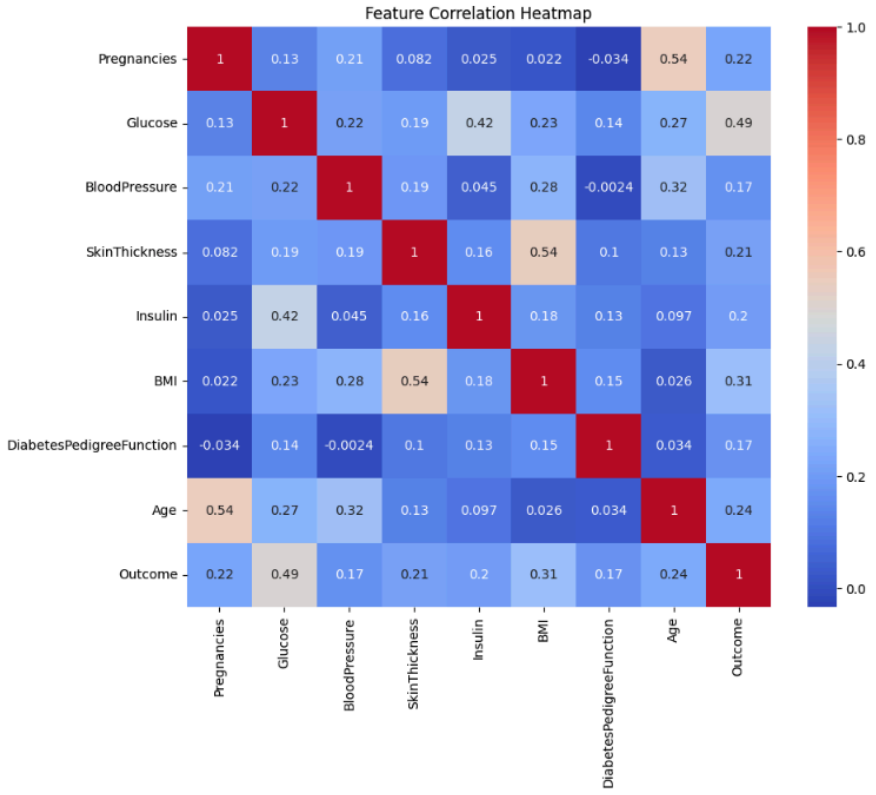


Based on the histograms in the above image and descriptive statistics, there are many different distributions. The distributions that have a right skew are pregnancies, insulin, DiabetesPedigreeFunction, and Age. The normal distributions are Glucose, Blood Pressure, and BMI. Outcome has a bimodal distribution with more at 0, indicating that the dataset is imbalanced and more patients don't have diabetes in the dataset than do have diabetes.



We can draw many conclusions from the boxplots given for each independent variable.

1. Pregnancies: Women with diabetes tend to have a higher median number of pregnancies compared to those without diabetes. Pregnant women with diabetes seem to have more variability.
2. Glucose: Glucose levels are significantly higher for women with diabetes compared to those without diabetes. This indicates that glucose is a good indicator of diabetes
3. Blood Pressure: Blood pressure appears similar across both groups, with comparable medians and ranges. However, there is more variability in individuals with diabetes
4. SkinThickness: Skin thickness is similar between the two groups, with slightly higher variability among diabetic individuals.
5. Insulin: Both diabetic and non-diabetic groups have a large range of insulin values and many outliers. Although medians are similar, individuals with diabetes have a slightly higher interquartile range, suggesting insulin levels contribute to some extent.
6. BMI (Body Mass Index): Women with diabetes generally have higher BMI values compared to non-diabetic individuals, with a clear upward shift in the median. This shows BMI's role in predicting diabetes outcomes.
7. Diabetes Pedigree Function: The diabetes pedigree function shows slightly higher values in women with diabetes, with greater variability and more outliers. This suggests that family history has a mild influence on diabetes outcomes.
8. Age: Diabetic individuals tend to be older on average, with a higher median age compared to non-diabetic individuals. Age seems to be an important factor in diabetes outcomes, as the likelihood of diabetes increases with age.



In this correlation heatmap, I tested how correlated various independent variables are with the outcome of being diabetic. The strongest correlation is with Glucose at 0.49, making it the most influential predictor. It indicates that higher glucose levels are more likely to have diabetes. The next strongest is BMI at 0.31 correlation, showing that a higher BMI is more likely to have diabetes. Lastly, age shows a smaller positive correlation at 0.24, showing that older people are slightly more likely to be diabetic. Other factors such as Pregnancy (0.22), (0.21), skin thickness (0.21), diabetes pedigree function (0.17), and blood pressure (0.17) exhibit weak correlation. Overall, glucose is the dominant factor, followed by BMI and Age, while other features exhibit limited direct influence on the Outcome variable.

Models and Methods

To predict whether or not a patient has diabetes, I decided to use multiple classification models and see which one performs the best in predicting diabetes outcomes and accounting for the variation in my data. For each of these models, I decided to utilize an 80-20 train-test split, training my models on 80% of the data and then testing them on the remaining 20%. The models I will use include Logistic Regression, Random Forest, and Gradient Boosting as the additional models to compare performance and assess predictive accuracy.
First, I evaluated the success of each of my models by comparing its performance metrics, such as the model's accuracy, against a baseline's accuracy. To get my baseline value, I simply predicted the most frequent class (either 0 or 1) from my dataset. Throughout this test we find that all the models performed better than the baseline test, indicating that they are worth the time and effort.

1) Logistic Regression

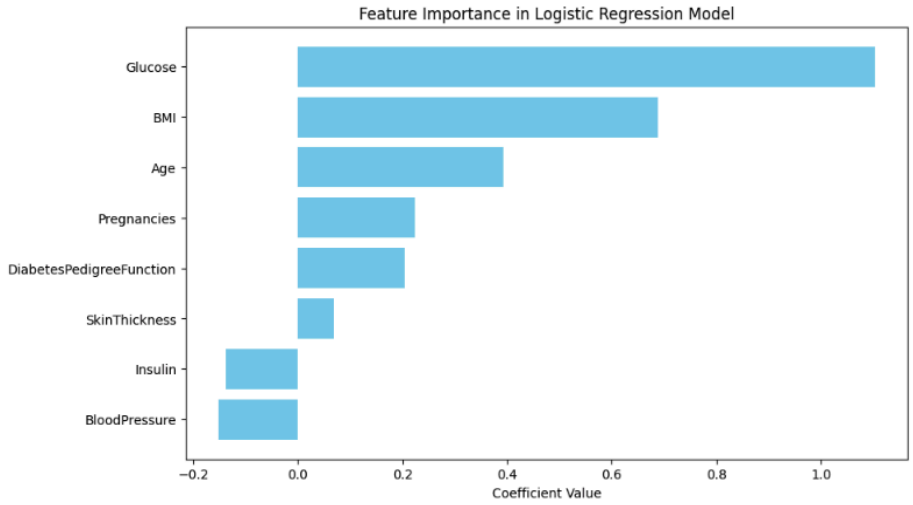
First, I chose to build a logistic regression model because I wanted to use independent variables to predict the dependent variable, as I believed these predictors may have collectively influenced the likelihood of diabetes. Logistic regression allowed me to model the relationships between the outcome (diabetes or no diabetes) and each of these predictors while also considering their combined effect.

Logistic Regression Performance:

	precision	recall	f1-score	support
0	0.80	0.83	0.81	99
1	0.67	0.62	0.64	55
accuracy			0.75	154
macro avg	0.73	0.72	0.73	154
weighted avg	0.75	0.75	0.75	154

The logistic regression model performed much better than the baseline model which had an accuracy of 0.65. I believe this is because the model could use the information contained in the independent variables to make predictions. The logistic regression model had a 0.75 accuracy, a 10% improvement from the baseline. For individuals with no diabetes, the model had a precision of 0.80, recall of 0.83, and an F1-score of 0.81, a relatively strong performance. For individuals with diabetes, the model performed moderately. It had a precision of 0.67, a recall of 0.62, and an F1 score of 0.64. Unlike the baseline model,, the logistic regression model could identify and correctly classify a significant portion of diabetes cases. The logistic regression model outperformed the baseline overall and could predict both classes, making it a better choice for this dataset.

I also examined feature importance to see what independent variables were most significant in the model.



Overall, the most significant features in predicting diabetes in the Logistic Regression Model were Glucose, BMI, and Age, which had coefficients ranging from 0.4-1.0. Features that had some impact were Age, Pregnancy, Diabetes Pedigree Function, and Skin Thickness, ranging from 0 -0.4. Insulin and blood pressure had minimal influence on the predictions made in the model, as seen by the negative coefficient value.

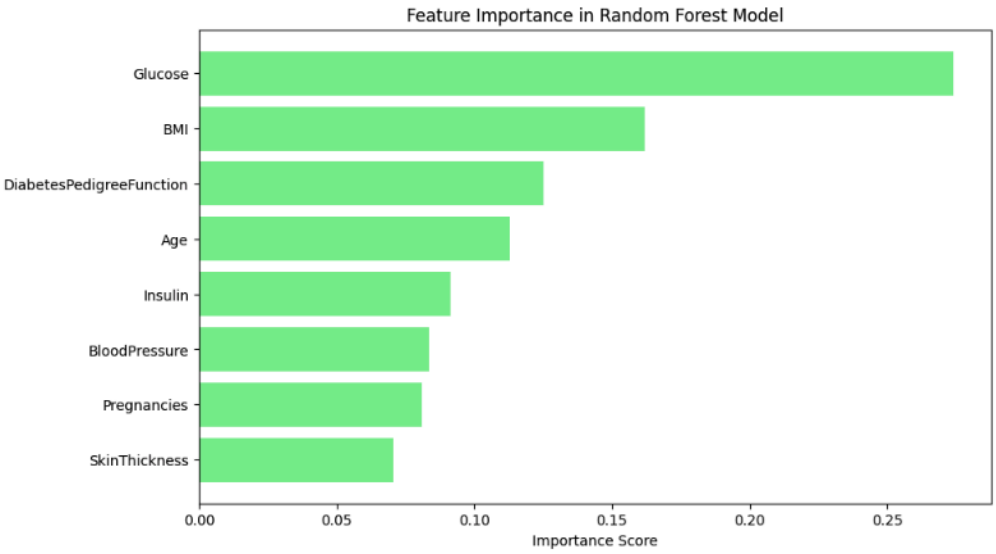
2) Random Forest Model

I also chose to build a Random Forest model because I wanted to use it to handle complex relationships between the predictors and the outcome (diabetes or no diabetes). Random Forest allowed me to assess the collective influence of multiple features, such as Glucose, BMI, and Age, on the likelihood of diabetes. Building decision trees means the model can capture intricate patterns in the data while reducing the risk of overfitting. Additionally, Random Forest tells us about feature importance, helping me identify which predictors contributed most significantly to the outcome.

```
Test Accuracy: 0.7337662337662337

Random Forest Report:
              precision    recall  f1-score   support

           0       0.76      0.87      0.81       101
           1       0.66      0.47      0.55        53

    accuracy                           0.73       154
   macro avg       0.71      0.67      0.68       154
weighted avg       0.72      0.73      0.72       154
```

Overall, my random forest model performed better than the baseline but worse than the logistic regression model. The random forest model achieved an accuracy of 73%, slightly worse than the logistic regression model. The random forest model had a better recall for class 0 at 0.87, showing it could identify

non-diabetic cases. However, the model struggled to identify diabetic cases, as shown by the 0.47 recall. In general, the logistic regression model had a higher precision, macro-average, weighted average, and f1 score than the random forest model, making it a better model for this dataset.



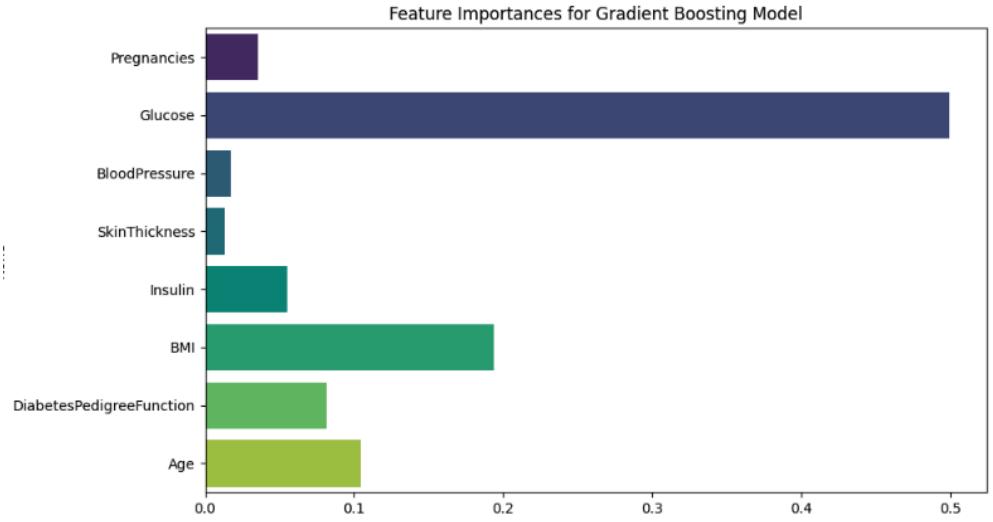
The most significant features in predicting diabetes in the Random Forest Model were Glucose and BMI, as they had the highest importance scores ranging from 0.15 -0.25. Features that had a moderate impact included Diabetes Pedigree Function, Age, and Insulin, which contributed to the model's predictions but to a lesser extent. Those ranged from 0.10-0.15. Blood Pressure, Pregnancy, and Skin Thickness showed relatively lower importance scores, indicating they had minimal influence on the predictions made in this model. Those ranged from 0 - 0.10.

3) Gradient Boosting

Lastly, I created a gradient-boosting model because it captures the complexity between the predictors and the outcome and optimizes the model's performance. Gradient Boosting sequentially combines decision trees to create a strong predictive model. Each new tree corrects the errors made by the previous ones. This iterative process helps improve accuracy as the model learns from the patterns. Additionally, Gradient Boosting helps us understand which features such as Glucose, BMI, and Age, had the most significant influence on the likelihood of diabetes.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.85      0.80       100
           1       0.64      0.50      0.56        54

    accuracy                           0.73       154
   macro avg       0.70      0.68      0.68       154
weighted avg       0.72      0.73      0.72       154
```

The gradient boosting report shows that the model is much better at predicting non-diabetic cases compared to diabetic cases. The precision, recall, and f1-score for diabetics is much lower, around the 0.5-0.7 range. Importantly, the recall is at 0.5, showing that the model only identifies half the diabetic cases. On the other hand, the precision, recall, and f1-score for non-diabetics is much higher, around 0.8 range.

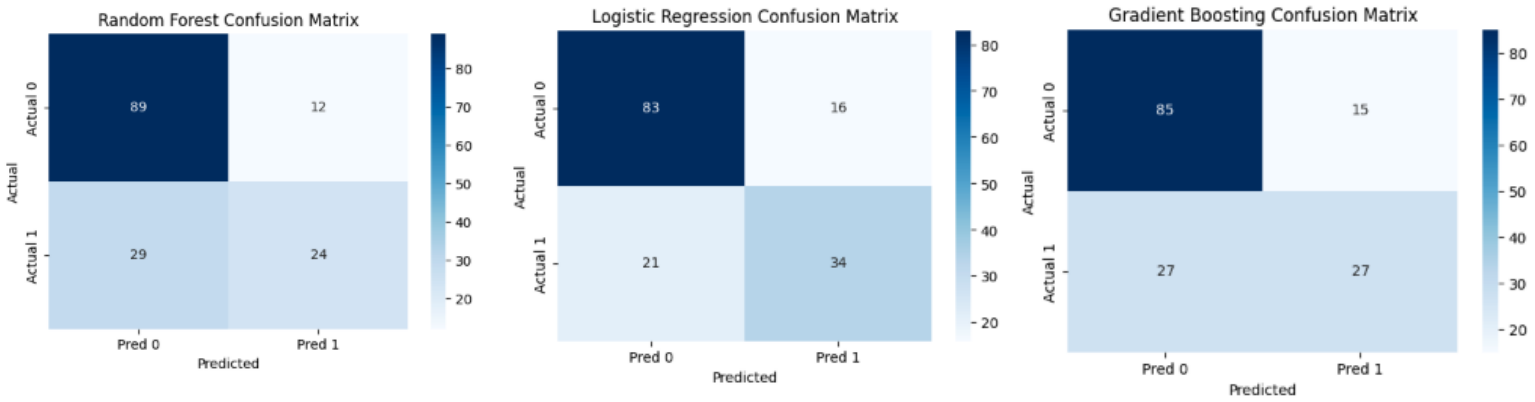


The most important feature by far in predicting diabetes in this Gradient Boosting Model was Glucose, with a value close to 0.5. BMI was the next largest impact, with an importance of 0.2. Other features such as Age, Diabetes, Pedigree Function, and Insulin had a medium to low impact on predictions. Lastly, Skin Thickness, Blood Pressure, and Pregnancy had minimal influence on how the model determined the outcome.

Result and Interpretation

Here I used confusion matrices to evaluate which model performed the best. The matrices are divided into 4 categories: true positives (correctly identified diabetes cases), false negatives (missed diabetes cases), true negatives (correctly identified no diabetes), and false positives (incorrectly identified diabetes). In medical cases, it's important to not have any false negatives. We would rather incorrectly identify diabetes than miss patients who actually have it. This is important when dealing with imbalanced classes because it tells us which model minimizes errors and effectively identifies positive cases of diabetes. Comparing

Confusion matrices allows us to see the strengths and weaknesses of each one and find the most reliable model that accurately predicts diabetes.



Key findings for the confusion matrices

1) The Random Forest Model Performed the Best: The Random Forest model demonstrates strong predictive performance. It achieved 24 true positives while only having 12 false positives and 29 false negatives. With 89 true negatives, the model is very good at classifying non-diabetes cases.

2) Performance of Gradient Boosting: The Gradient Boosting model performed slightly worse than Random Forest but still had good predictive power. It captured 27 true positives and produced 15 false positives and 27 false negatives. While the false negatives remain relatively high, the performance is better balanced than Random Forest in identifying class 1 (diabetic cases), and its 85 true negatives indicate that it is very good at detecting someone who doesn't have diabetes.

3) Logistic Regression Struggled with Identifying Diabetes: The Logistic Regression model had more difficulty identifying diabetes cases. It achieved 34 true positives, which is the highest among the three models, but also had 21 false negatives. The number of false positives (16) remains moderate, and the model performed well in predicting non-diabetes cases with 83 true negatives. Its higher recall for diabetes cases suggests it is better for scenarios when minimizing missed diagnoses (false negatives) is more important.

Next Steps & Discussion

Summary of Findings

In my analysis of diabetes prediction, all the models demonstrated improved performance over the baseline predictor, confirming their usefulness in identifying positive cases (diabetes). The models, ranked in order of performance, are as follows: Logistic Regression, Gradient Boosting, and Random Forest

Key findings:

1) The success of the Logistic Regression Model: The Logistic Regression model was the most effective, showcasing the best predictive capabilities out of all the models. It is the best suited to capture the complex relationships in the diabetes data.
2) Impactful features: As seen throughout all three models, the Glucose level and the BMI of an individual were the best metrics to predict if they had diabetes in all the models. Age had a moderate also had a moderate impact on all of the models.
3) Variable Influence: In contrast, metrics such as Insulin, Blood pressure, Diabetes Pedigree Function, Skin Thickness, and pregnancies had a lower impact on predicting if one had diabetes, and their significance varied among the 3 models. Even though they are less influential, they still provided insights into models that made them more accurate when they were tested.

Next Steps/Improvements

To improve the predictive capabilities of the models and gain deeper insights into diabetes diagnosis, I would want to incorporate these additional features into my models.

1. Understanding Type 1 vs Type 2 Diabetes
    - There are different indicators for Type 1 and Type 2 diabetes, so knowing what type a patient has means we can look at different indicators. If we know to look at different indicators, we can predict the outcome more accurately and quickly
2. Dietary and Lifestyle Information
    - Lifestyle is a huge factor that can cause someone to have diabetes. As such, including data on dietary habits, physical activity levels, and smoking or alcohol consumption could provide good information about how lifestyle factors influence diabetes outcomes.
3. Patient Medical History:
    - Knowing what other medical conditions an individual has can help us better predict whether they are more likely to have diabetes.
4. Socioeconomic and Demographic Factors
    - Including features such as income levels, education, employment status, and access to healthcare could help understand how socioeconomic factors influence diabetes diagnosis. This is because these factors can influence an individual's access to medical care, healthy foods, and fitness resources, which could impact how common diabetes is for various groups

By integrating these additional factors into the analysis, I can refine the models more and get a better understanding of the factors that contribute to whether someone is diabetic. This could lead to even more accurate predictions for healthcare professionals and patients to use.