Predictive Factors for Diabetes in Adults

By Tom Herman
STAT E-109, Spring 2025
May 4, 2025

Research Question

This project investigates the question: ***What factors are most predictive of diabetes risk in adults, and how accurately can we model and predict that risk using statistical methods?*** By using logistic regression and random forest machine learning algorithms, I seek to uncover which health markers and demographic features (such as glucose levels, BMI, insulin levels and age) contribute most strongly to the likelihood of developing diabetes, while identifying the best algorithm to predict the presence of diabetes.

Motivation

I have personally witnessed the devastating effects of both Type I and Type II diabetes on my loved ones, including the recent and untimely deaths of my mother and brother-in-law. These losses have deepened my awareness of how serious and widespread this disease can be. I chose this dataset to explore the relationship between common health markers and the likelihood of developing diabetes, not only to gain insight into the most significant risk factors, but also to inform my own health choices and potentially help others avoid similar outcomes.

Hypotheses

Based on my own basic understanding of diabetes and its risk factors, this project proposes two main hypotheses:

1. Among the variables provided in this dataset, **glucose and insulin levels will be the most significant predictors** of diabetes. Elevated glucose levels are a direct diagnostic marker of diabetes, while abnormal insulin levels often reflect insulin resistance (a key indicator of diabetes or pre-diabetes).
2. **A predictive model** (using logistic regression and/or random forest) can be trained on the data to **reliably estimate an individual's likelihood of having diabetes**.

Assumptions

This analysis is built on several important assumptions, especially for the logistic regression model. One key assumption is that there's a **linear relationship** between each predictor and the log-odds of the outcome (in this case, whether or not someone has diabetes). It also assumes that each observation in the dataset is **independent**, meaning no individual's data is directly tied to another's. I assume that is the case here since there is no information about how the data were actually collected.

Methods

To explore the factors most strongly linked to diabetes and build a predictive model, I used the **“SEMMA” data mining workflow** (Sample, Explore, Modify, Model, Assess) to create two different predictive models: logistic regression and random forest. Logistic regression is helpful for understanding which individual features—like glucose or insulin—are associated with increased diabetes risk. Random forest, on the other hand, is a powerful machine learning algorithm that can capture more complex, non-linear relationships and often performs better at prediction. I trained and tested both models to compare how well they performed using common evaluation metrics like accuracy, sensitivity (true positive rate), specificity (true negative rate) and the area under the ROC curve (AUC).

The data for this project comes from the **Healthcare Diabetes Dataset**, which I downloaded from Kaggle (see Reference 1). It includes health information such as number of pregnancies, glucose levels, BMI, insulin, blood pressure, and age for each individual, along with a label indicating whether or not they have diabetes. These features are commonly used in diabetes screening and make the dataset well suited for both analysis and prediction. The complete list of all 10 variables is shown below:

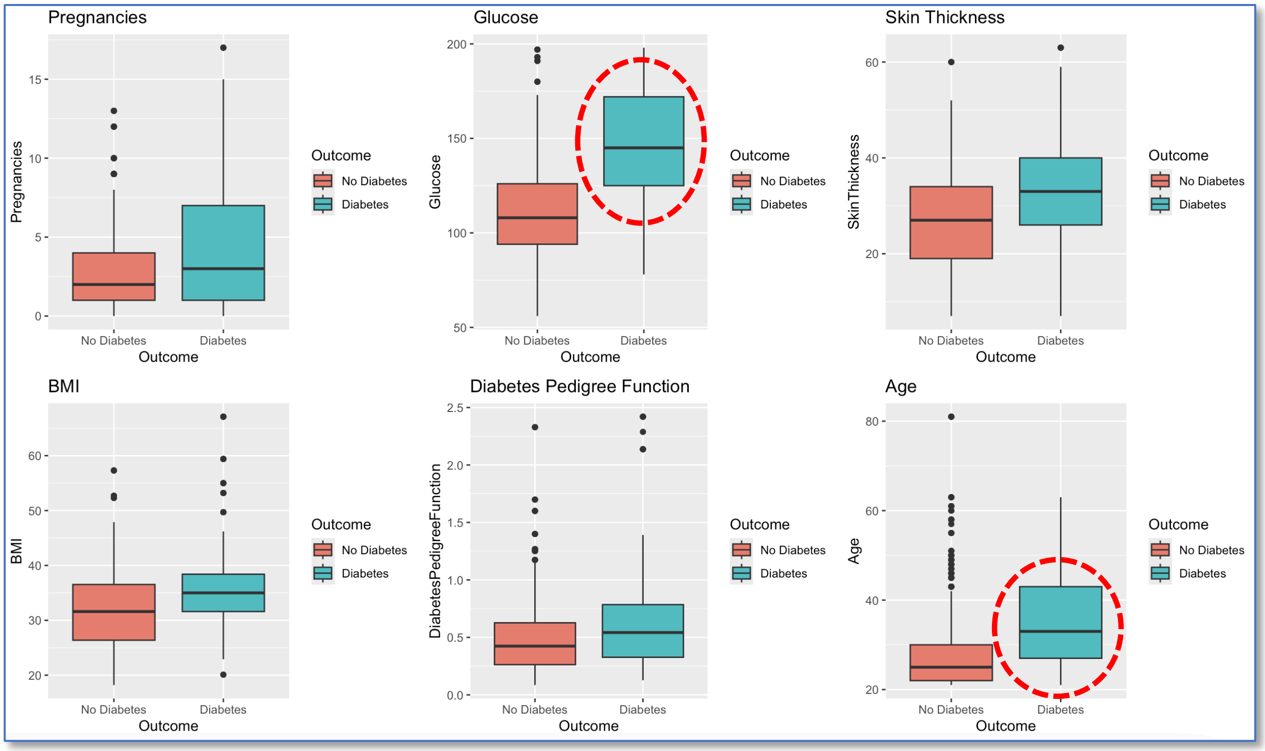
Variable	**Description**
Id	Unique ID for each data entry.
Pregnancies	Number of times pregnant.
Glucose	Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
BloodPressure	Diastolic blood pressure (mm Hg).
SkinThickness	Triceps skinfold thickness (mm).
Insulin	2-hour serum insulin (mu U/ml).
BMI	Body mass index (weight in kg/height in m^2).
DiabetesPedigreeFunction	Diabetes pedigree function, a genetic score of diabetes.
Age	Age in years.
Outcome (Predicted Variable)	Binary classification indicating the presence (1) or absence (0) of diabetes.

Before modeling, I **cleaned the dataset** by replacing zero values (which are biologically implausible) with NA, and then removed rows with missing values. This reduced the dataset from 2,768 to 1,427 complete records. I also evaluated the data for outliers, but did not remove any potential outliers due to my own limited knowledge around the plausibility of those outliers (people’s physical health and related markers can vary widely). See the “Limitations” section for more detail about these issues and their impacts.

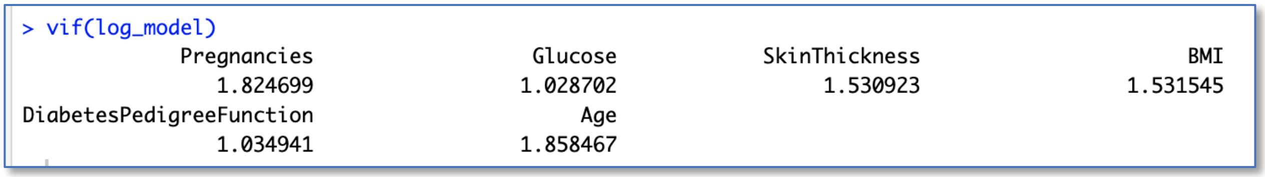
After cleaning the dataset, I **partitioned** the cleaned data by sampling 70% into a dataset to train the models with, and 30% to actually test the models and assess their accuracy.

To get an initial sense of the data and whether my first hypothesis is correct, I performed some preliminary **Exploratory Data Analysis (EDA)**. The boxplots below show how distributions of each variable and how they differ between people with and without diabetes.

Right off the bat, I found a big disparity between people with and without diabetes by the Glucose and Age variables, where people with diabetes tend to have much higher glucose levels and also tend to be older. These two variables have nearly completely different interquartile ranges for "no diabetes" vs. "diabetes." This led me to suspect that the models will identify these two variables as some of the strongest predictors of diabetes.



Finally, to reduce the risk of **multicollinearity** (when predictors are too closely related), I reviewed correlations and confirmed that no major issues were present. The table below shows the Variance Inflation Factor (VIF) values for the six most significant variables:



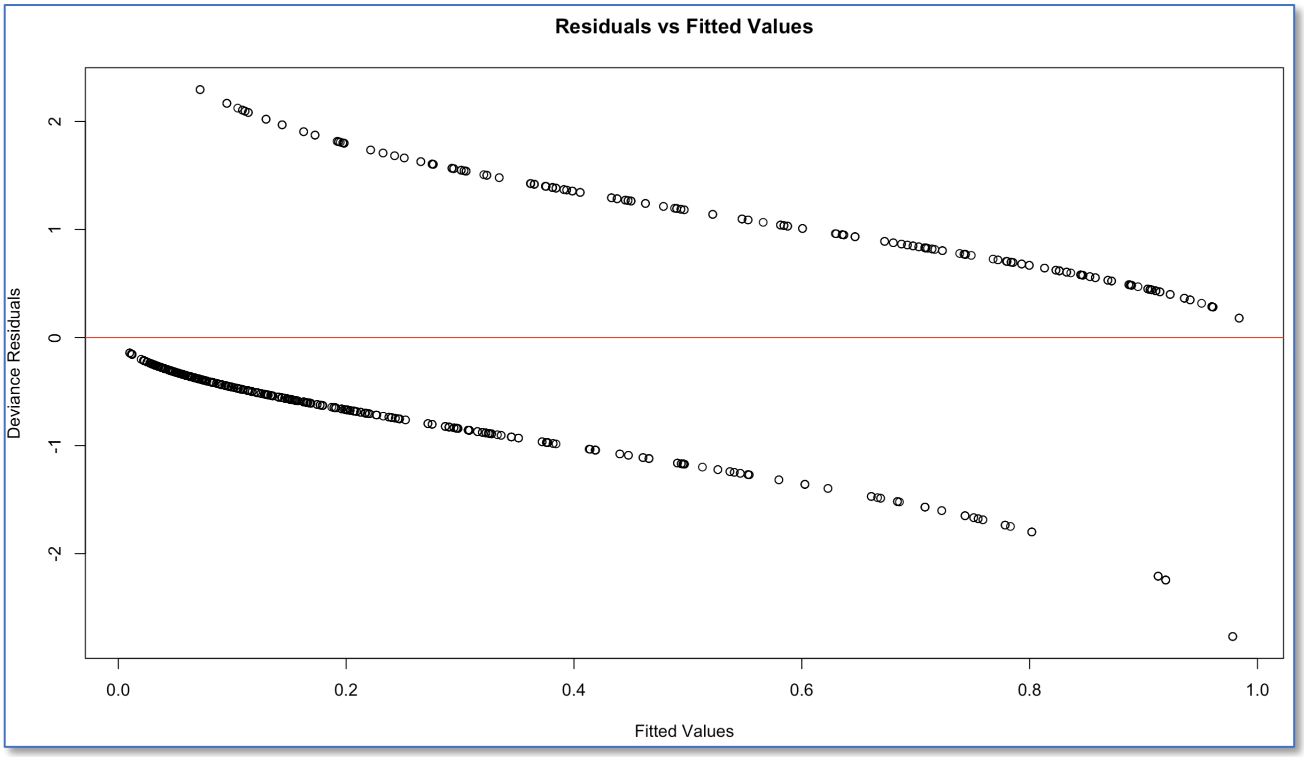
```
> vif(log_model)
             Pregnancies                  Glucose            SkinThickness                      BMI
                1.824699                 1.028702                 1.530923                 1.531545
DiabetesPedigreeFunction                      Age
                1.034941                 1.858467
```

All of the VIF values are between 1 and 2, which indicates very low multicollinearity.

Results

After running the logistic regression model, I noticed it reported two variables with higher P-values (suggesting these variables were not influential): Blood Pressure and Insulin. I re-ran the model with those variables excluded, but doing so didn't make much of a difference in the model's accuracy.

While my final logistic regression model showed decent accuracy and consistency on the training and test data (78.9% and 78.3%, respectively), there was a clear pattern in the Residuals vs. Fitted Values plot (shown below), which suggests one of my initial assumptions (linear relationship) is incorrect, and the actual relationships may be non-linear or more complex.



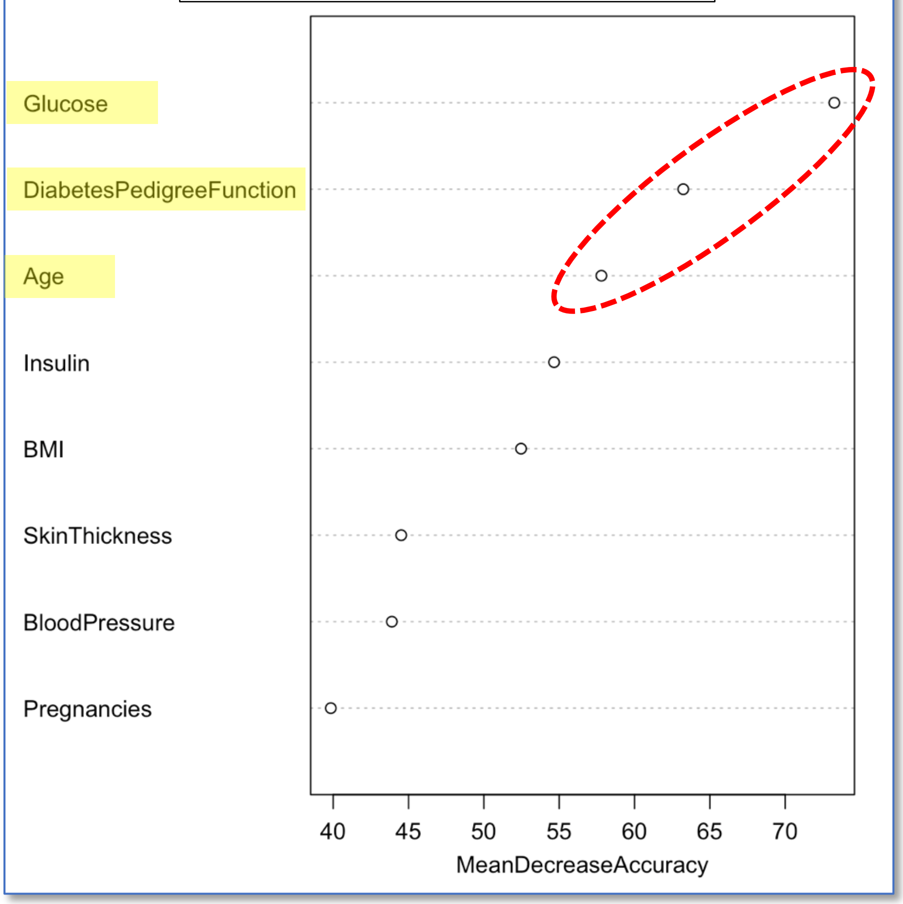
After running the random forest model, it proved to be **significantly more accurate and much better suited** to train and make predictions on this dataset. The model accuracy on testing data went from 78.3% for the logistic regression model up to 98.1% for the random forest model. The random forest's accuracy on the training data was 100%, which could suggest it is overfit. However, the model's very strong performance on the unseen data (test dataset) makes this less of a concern. Continuing on the following page is a comprehensive comparison of both models and their performances on the test data.

According to **sensitivity and specificity**, the random forest is also much stronger at detecting actual diabetes cases and correctly identifying non-diabetic individuals. The latter is important because I want to avoid false positives (false alarms). The higher **F1 Score** (0.97) reflects the balance of both false positives and false negatives, which is significantly improved for the random forest model compared to the logistic regression model. **AUC** (area under the curve) measures the model's ability to distinguish between classes. While the logistic regression AUC of 0.84 is very good, random forest is nearly perfect (0.99).

Model Performance Comparison on Test Data		
Metric	**Logistic Regression**	**Random Forest**
Accuracy	78.3%	98.1%
Sensitivity	56.7%	96.5%
Specificity	88.9%	98.9%
Kappa	0.48	0.96
F1 Score	0.63	0.97
AUC	0.84	0.99

Finally, I used the **variable importance** function on the random forest model to determine the most influential variables in the prediction of diabetes. The chart to the right shows the importance of each variable, from most influential at the top, to least influential at the bottom. As suspected in my original hypothesis, **glucose level** proved to be the most significant variable. To my surprise however, the "Diabetes Pedigree Function" and Age turned out to be the second- and third-most influential variables. Insulin level was less influential, at fourth down the list. BMI was similar in influence to Insulin, and Skin Thickness, Blood Pressure and Pregnancies were the least influential.

Random Forest Variable Importance



In summary, while the logistic regression model offered reasonable accuracy and interpretability, it showed limitations in capturing the complex, potentially non-linear relationships in the data. The random forest model, on the other hand, demonstrated exceptional performance across nearly all metrics, with higher accuracy, stronger agreement with actual outcomes, and better balance between sensitivity and specificity. The model's ability to rank variable importance also helped validate key predictors, such as glucose, age, and the Diabetes Pedigree Function, offering deeper insight into what drives diabetes predictions in this dataset.

Conclusion

The original motivation for this project came from a personal place: a desire to better understand the risk factors for diabetes, a disease that has affected people close to me. Going into this analysis, I hypothesized that certain variables (particularly glucose level and insulin) would be the strongest predictors of diabetes. I also wanted to explore whether a more complex model like random forest would outperform a traditional logistic regression approach.

After cleaning the data and conducting exploratory data analysis, I noticed some variables, like glucose and age, showed noticeably different distributions between those with and without diabetes. In checking for outliers, I found a few unusual values but decided to keep them in the analysis due to the possibility that they were real and plausible. I also made necessary assumptions (like independence of observations) based on the limitations of the dataset and acknowledged those where appropriate.

The logistic regression model performed fairly well, achieving test accuracy around 78%. However, the residual diagnostics indicated potential issues with linearity, suggesting that the model might be too simple to fully capture the relationships in the data. In contrast, the random forest model produced excellent results across all key metrics, with nearly perfect test accuracy (98%), a high F1 Score (0.97), and an AUC of 0.99. It also confirmed that glucose is the most influential predictor, followed by Diabetes Pedigree Function and age (both of which I hadn’t initially expected to be quite so important). Interestingly, insulin was not as predictive as I had assumed at the outset, which challenged one of my early beliefs.

In the end, the analysis supported my core hypothesis that glucose level is a key driver of diabetes risk, but it also broadened my understanding of how other variables, like family history and age, factor in. From a modeling perspective, this project reinforced the idea that more flexible, non-linear models like random forest can often better capture the complexity in real-world health data. It also highlighted the tradeoffs between interpretability and performance when choosing between statistical and machine learning models.

This project gave me both technical experience in building and evaluating models, and personal insight into what drives this unfortunate disease. While the models aren’t perfect and the dataset had its limitations, I feel this analysis offered a meaningful step toward better understanding the patterns and predictors of diabetes.

Potential Limitations and Challenges

Data Loss

The dataset included some implausible values of “0” for measurements like Glucose, Blood Pressure, Skin Thickness, Insulin and BMI. In other words, if the value of any of these measurements was truly 0, the person would be either deceased, or experiencing a medical emergency. Therefore, those values were identified and replaced with N/A.

After removing all rows with missing data (N/A values), the dataset decreased from 2,768 rows to 1,427 rows. While a 48% reduction is not ideal, it still provides enough data to train and test models with. The cleaned data remained balanced, retaining roughly an equity proportion of both classes of 0 (no diabetes) vs. 1 (diabetes).

Independence Assumption

Another limitation is that the author of the dataset did not provide any detail on how the data were collected, so it is unclear whether any observations may be related (for example, from the same family or clinic). My analysis assumes independence of observations, but if that assumption is violated, it could affect the validity of the results.

Outliers

As part of my data preparation, I checked for outliers in all numeric variables by looking at interquartile ranges. Several variables (particularly Insulin, Age, and the Diabetes Pedigree Function) showed some values that fell outside the typical range. While some of these may represent unusual or extreme cases, they could also be valid observations, especially given the variability of health data across individuals. Since I don’t have the expertise to confidently determine which of these values are legitimate and which might be errors, I kept all outliers in the dataset.

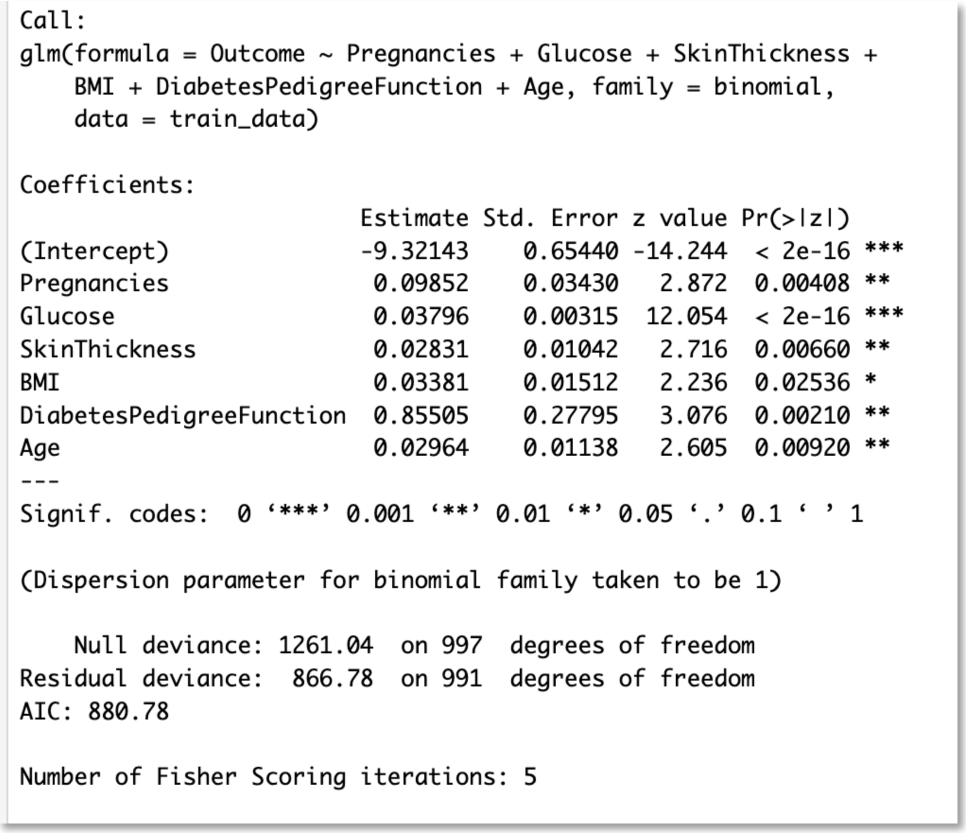
References

Nandita Pore
Healthcare Diabetes Dataset: A Comprehensive Dataset for Diabetes Risk Assessment
Kaggle: https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes
Accessed May 2025

Appendix

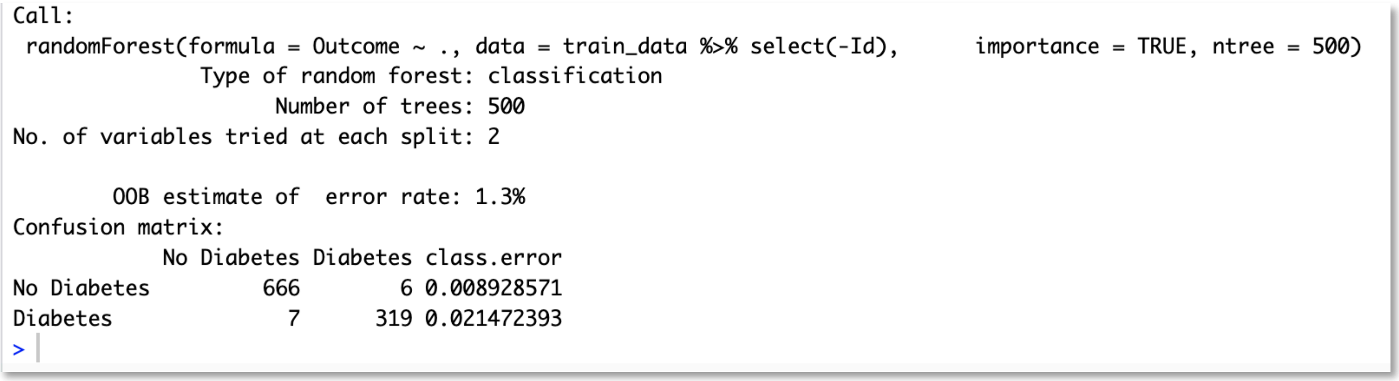
On the following pages are some important outputs from the analysis performed using R, including summary outputs of the logistic regression and random forest models, as well as the calculation of important model performance metrics.

Result of GLM Model Summary on Training Data:



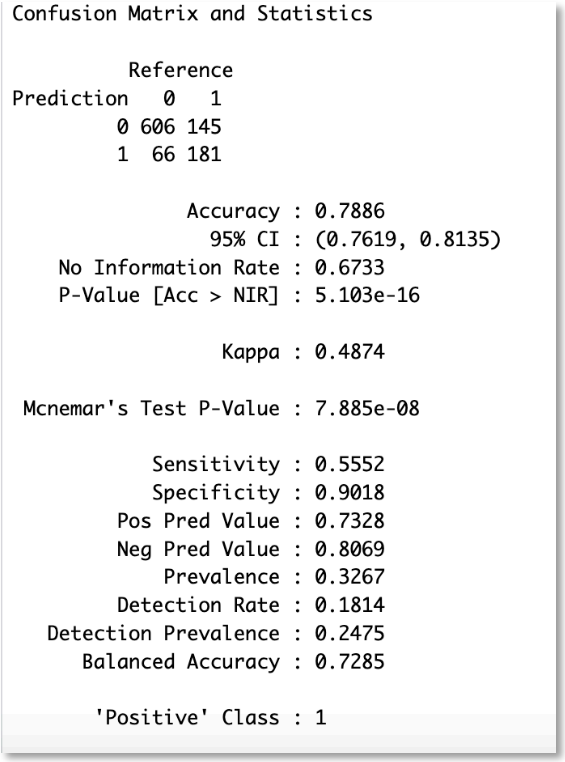
```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + SkinThickness +
    BMI + DiabetesPedigreeFunction + Age, family = binomial,
    data = train_data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.32143    0.65440 -14.244  < 2e-16 ***
Pregnancies               0.09852    0.03430   2.872  0.00408 **
Glucose                   0.03796    0.00315  12.054  < 2e-16 ***
SkinThickness             0.02831    0.01042   2.716  0.00660 **
BMI                       0.03381    0.01512   2.236  0.02536 *
DiabetesPedigreeFunction  0.85505    0.27795   3.076  0.00210 **
Age                       0.02964    0.01138   2.605  0.00920 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1261.04  on 997  degrees of freedom
Residual deviance:  866.78  on 991  degrees of freedom
AIC: 880.78

Number of Fisher Scoring iterations: 5
```

Result of Random Forest Model Summary on Training Data:



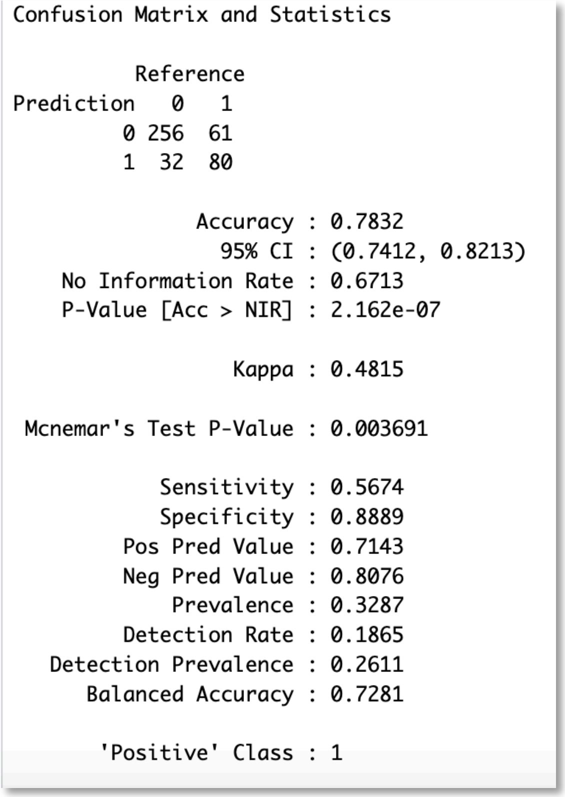
```
Call:
 randomForest(formula = Outcome ~ ., data = train_data %>% select(-Id),      importance = TRUE, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 1.3%
Confusion matrix:
            No Diabetes Diabetes class.error
No Diabetes         666        6 0.008928571
Diabetes              7      319 0.021472393
> 
```

Result of GLM Model Confusion Matrix (Training Data):



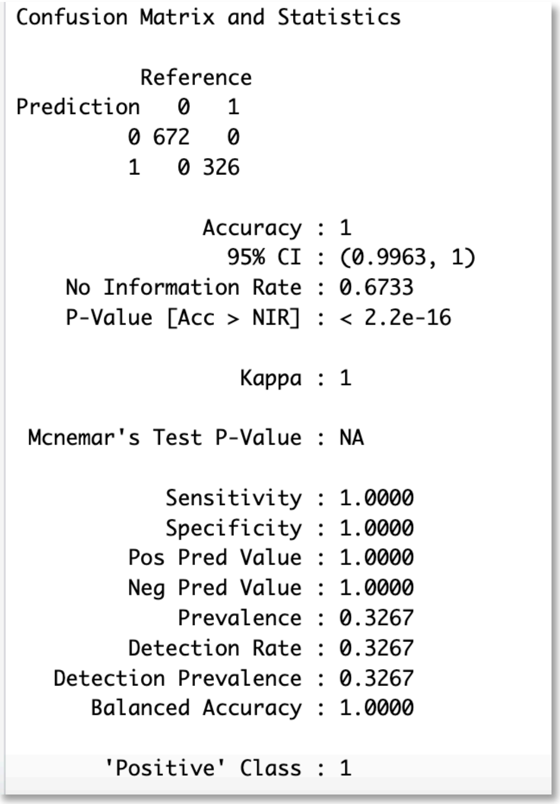
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 606 145
         1  66 181

               Accuracy : 0.7886
                 95% CI : (0.7619, 0.8135)
    No Information Rate : 0.6733
    P-Value [Acc > NIR] : 5.103e-16

                  Kappa : 0.4874

 Mcnemar's Test P-Value : 7.885e-08

            Sensitivity : 0.5552
            Specificity : 0.9018
         Pos Pred Value : 0.7328
         Neg Pred Value : 0.8069
             Prevalence : 0.3267
         Detection Rate : 0.1814
   Detection Prevalence : 0.2475
      Balanced Accuracy : 0.7285

       'Positive' Class : 1
```

Result of GLM Model Confusion Matrix (Test Data):



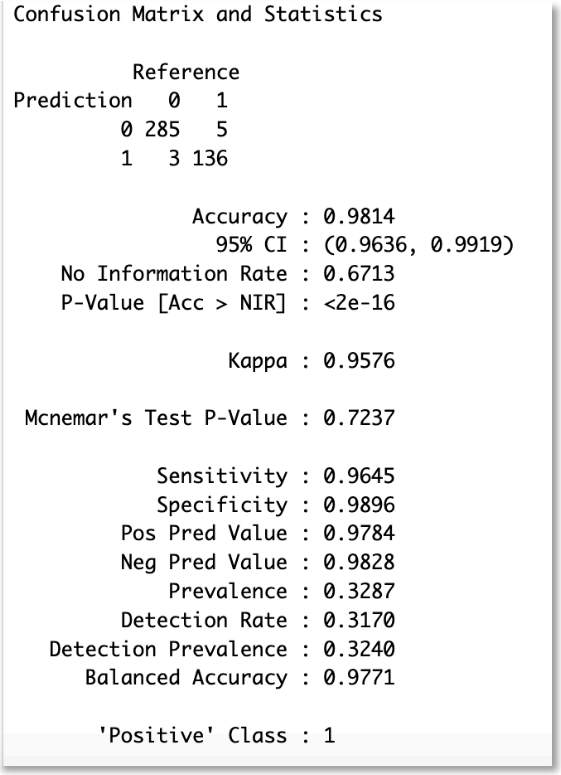
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 256  61
         1  32  80

               Accuracy : 0.7832
                 95% CI : (0.7412, 0.8213)
    No Information Rate : 0.6713
    P-Value [Acc > NIR] : 2.162e-07

                  Kappa : 0.4815

 Mcnemar's Test P-Value : 0.003691

            Sensitivity : 0.5674
            Specificity : 0.8889
         Pos Pred Value : 0.7143
         Neg Pred Value : 0.8076
             Prevalence : 0.3287
         Detection Rate : 0.1865
   Detection Prevalence : 0.2611
      Balanced Accuracy : 0.7281

       'Positive' Class : 1
```

Result of Random Forest Model Confusion Matrix (Training Data):



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 672   0
         1   0 326

               Accuracy : 1
                 95% CI : (0.9963, 1)
    No Information Rate : 0.6733
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.3267
         Detection Rate : 0.3267
   Detection Prevalence : 0.3267
      Balanced Accuracy : 1.0000

       'Positive' Class : 1
```

Result of Random Forest Model Confusion Matrix (Test Data):



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 285   5
         1   3 136

               Accuracy : 0.9814
                 95% CI : (0.9636, 0.9919)
    No Information Rate : 0.6713
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9576

 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.9645
            Specificity : 0.9896
         Pos Pred Value : 0.9784
         Neg Pred Value : 0.9828
             Prevalence : 0.3287
         Detection Rate : 0.3170
   Detection Prevalence : 0.3240
      Balanced Accuracy : 0.9771

       'Positive' Class : 1
```