Jeff Nania
BFOR 516
Final Project

Data

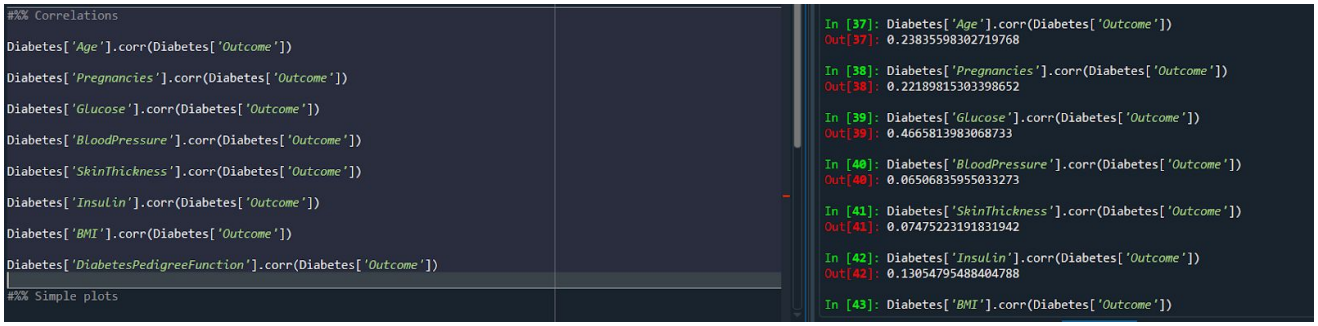
The data for this project comes from UCI Machine Learning, but is available at Kaggle through the following link: https://www.kaggle.com/uciml/pima-indians-diabetes-database . The data is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The patients considered in this dataset are all females above the age of 21 of Pima Indian Heritage. The Pima people are a group of Native Americans who live in what is now south and central Arizona as well as northwestern Mexico.

This data set considers factors such as glucose concentration, blood pressure, body mass index, age, and more to attempt to predict whether or not an individual has diabetes.

Questions

1. A. What factor or factors seem to be most closely linked with diabetes?

   By using simple correlations, we can see that Glucose and BMI have the highest correlation with Diabetes. Glucose has a score of roughly 0.47 and BMI has a score of roughly 0.29



```
#%% Correlations

Diabetes['Age'].corr(Diabetes['Outcome'])

Diabetes['Pregnancies'].corr(Diabetes['Outcome'])

Diabetes['Glucose'].corr(Diabetes['Outcome'])

Diabetes['BloodPressure'].corr(Diabetes['Outcome'])

Diabetes['SkinThickness'].corr(Diabetes['Outcome'])

Diabetes['Insulin'].corr(Diabetes['Outcome'])

Diabetes['BMI'].corr(Diabetes['Outcome'])

Diabetes['DiabetesPedigreeFunction'].corr(Diabetes['Outcome'])

#%% Simple plots
```

```
In [37]: Diabetes['Age'].corr(Diabetes['Outcome'])
Out[37]: 0.23835598302719768

In [38]: Diabetes['Pregnancies'].corr(Diabetes['Outcome'])
Out[38]: 0.22189815303398652

In [39]: Diabetes['Glucose'].corr(Diabetes['Outcome'])
Out[39]: 0.4665813983068733

In [40]: Diabetes['BloodPressure'].corr(Diabetes['Outcome'])
Out[40]: 0.06506835955033273

In [41]: Diabetes['SkinThickness'].corr(Diabetes['Outcome'])
Out[41]: 0.07475223191831942

In [42]: Diabetes['Insulin'].corr(Diabetes['Outcome'])
Out[42]: 0.13054795488404788

In [43]: Diabetes['BMI'].corr(Diabetes['Outcome'])
```

   B. What factor or factors seem to have the least correlation with diabetes?

   Also using simple correlations, we can see that Skin Thickness and Blood Pressure have the least correlation with diabetes. Skin Thickness has a score of roughly 0.075, and Blood Pressure has a score of 0.065.

C. What are the average Glucose and BMI of individuals with and without diabetes?

The average glucose for individuals with diabetes is 141.257 while the average glucose for individuals without diabetes was 109.98. The average BMI for individuals with diabetes was 35.1425, and the average BMI for individuals without diabetes was 30.3042.

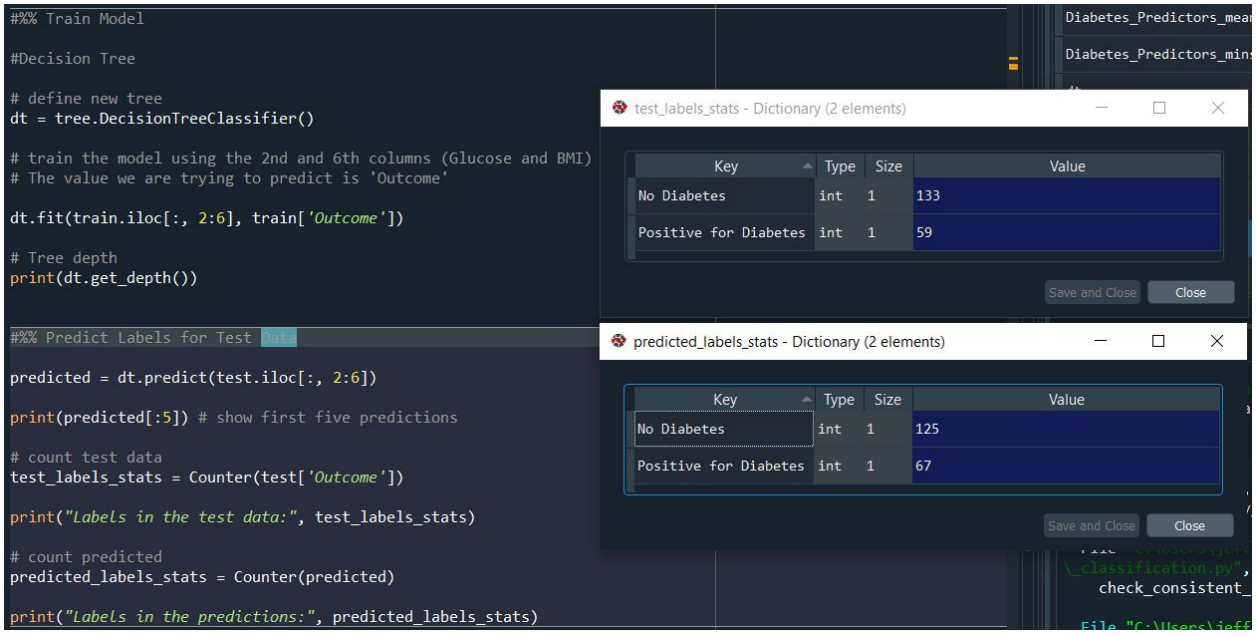
Diabetes_Predictors_means - DataFrame

Outcome	regnancie	Glucose	BloodPressure	SkinThickness	Insulin	BMI	sPedigreel	Age
0	3.298	109.98	68.184	19.664	68.792	30.3042	0.429734	31.19
1	4.86567	141.257	70.8246	22.1642	100.336	35.1425	0.5505	37.0672

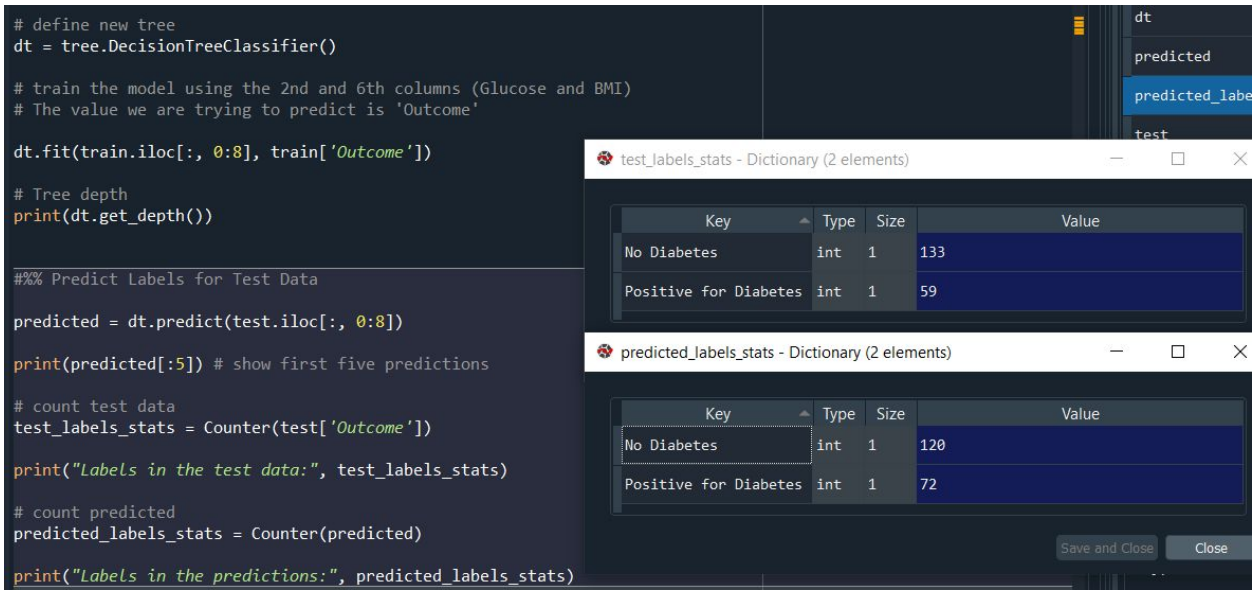
D. What is the breakdown of individuals with diabetes vs without diabetes for this dataset?

```
In [44]: print(Diabetes.Outcome.value_counts())
No Diabetes                500
Positive for Diabetes      268
Name: Outcome, dtype: int64
```

2. A. Split data set into train and test, and implement a decision tree using just those variables with the highest correlation to diabetes (glucose, BMI).

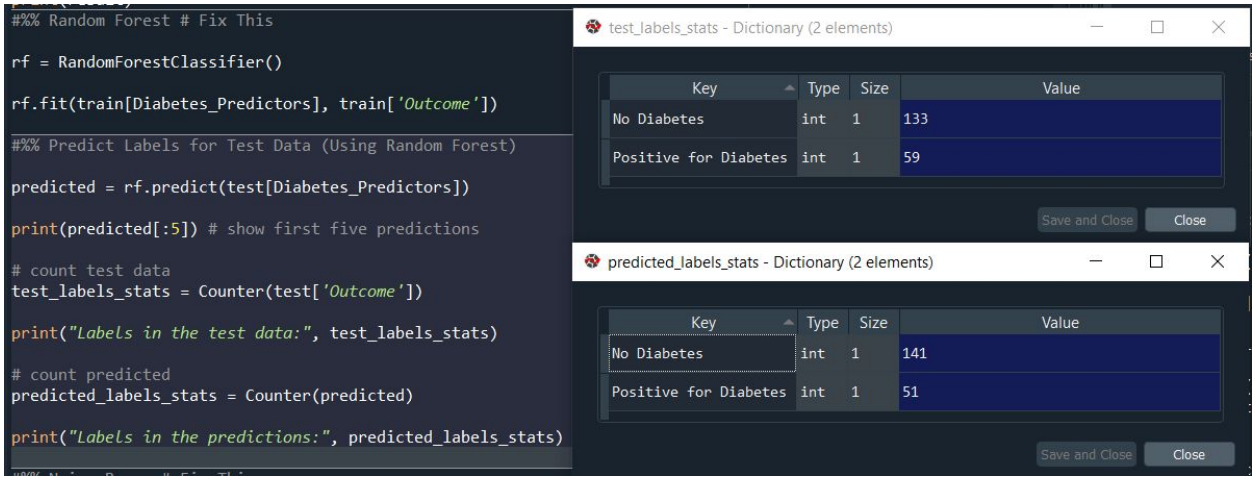


B. Run the same train and test set but this time implement a decision tree with all possible variables. How does the output change?



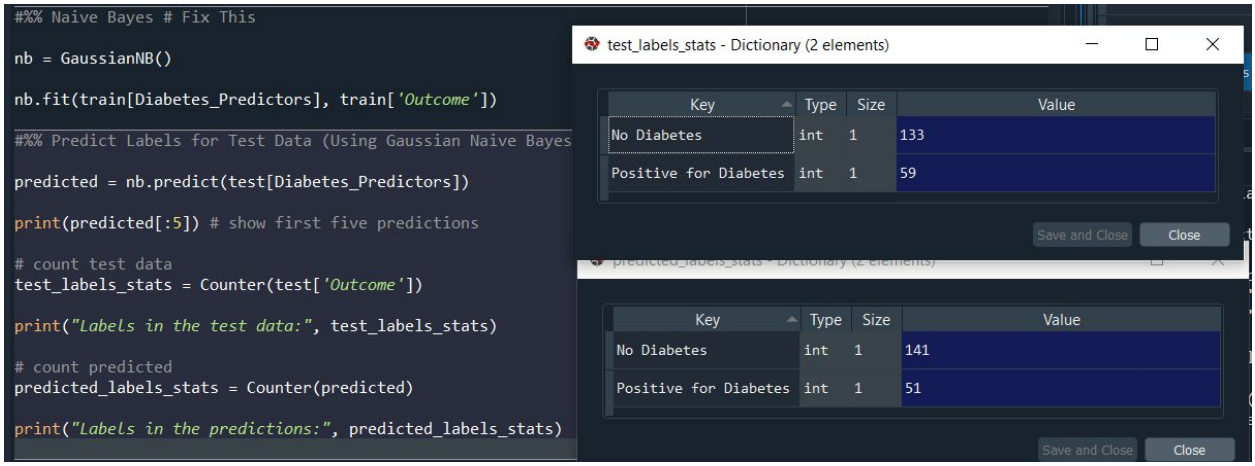
First of all -- the test label stats remain the same as they should, but the predicted stats went down to 120 for No Diabetes, and predicted 72 positive for diabetes. In this way the decision tree model was more accurate when only considering the two most highly correlated variables as it predicted 125 for No Diabetes which is closer to the test stat of 133.

C. Now implement a Random Forest Classifier which considers all the variables. How does it compare to the previous model?



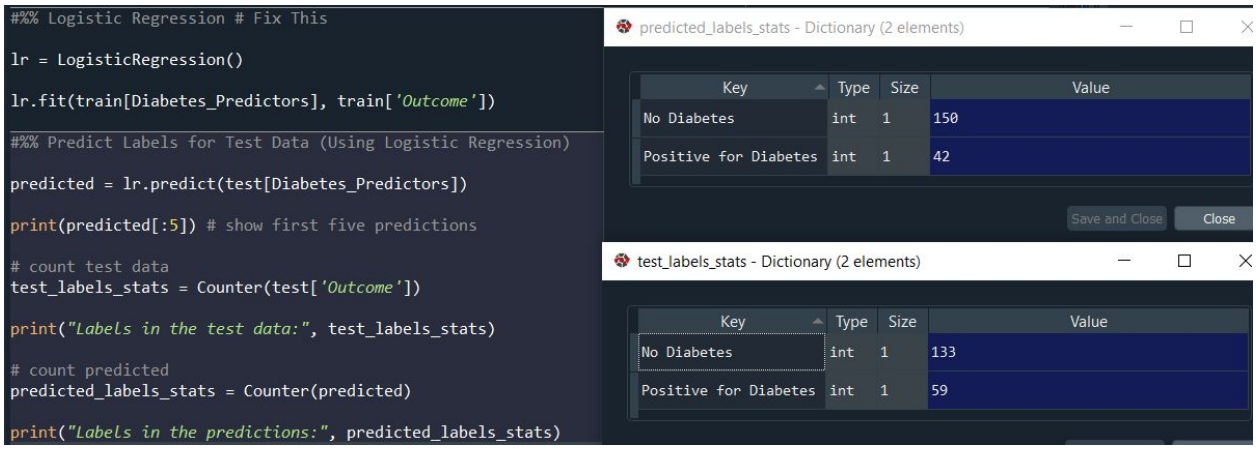
As we can see here this model significantly overshot the mark by predicting 141 with no diabetes as opposed to the two decision tree models which predicted 120 and 125 respectively.

D. Do the same as above but implement a Gaussian Naive Bayes model.



We can see here that this model also significantly overshot its prediction for No Diabetes. This model is the least accurate so far.

E. Do the same as above but implement a Logistic Regression model.



This one is the worst yet with significantly higher false prediction for No Diabetes.

3. Which models seem to perform the best for this prediction?

   Out of all the models, using a decision tree with only the two most correlated variables seems to work the best.