Muhammad Abdullah Umar, Netid: mumar4

Joshua Herman, Netid: jherma20

Hamza Shahid, Netid: hshahi2

**418 Final Project Write-up: Analysis of Pima Indians and Diabetes**

***1. Problem Selection. Identify a real-world problem (for example, predicting the number of votes cast for Democratic and Republican candidates in every U.S. county during the 2016 Presidential primaries) and propose a data science solution (for example, building linear regression models). Describe your problem and your solution.***

**The problem**: Predicting the outcome of type-2 diabetes (1) or no diabetes (0) for females for the Native American Tribe Pima Indians. The tribe has a statistically much higher rate of diabetes compared to the average American. Among Americans, 9.4% of the population has diabetes, and 25.82% have prediabetes which means they are at a high risk for type-2 diabetes.1 In this dataset of this tribe we have 65.1% of the women have diabetes. They are also a very unhealthy population with BMI’s averaging in the obese range, while the average American is only in the overweight range.

The Pima Native American Indians have been suffering from one of the highest rates of diabetes in the world, and this has intrigued researchers for years. Given the high rate of diabetes in Pima Native American Indians, they suffer from poor health conditions.2 In the USA this dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**Solution:** Exploring the diagnostic variables and making single and multiple linear regression to find statistically significant variables. Afterwards, classifying data based on the significant variables to find the highest contributing factors and the best predictors to diabetes for the Pima Indians.

**The dataset:**  (https://www.kaggle.com/uciml/pima-indians-diabetes-database/home)

***2. Data Collection. Identify one or more datasets relevant to your problem. Describe your datasets.***

**Shape:** 768 rows × 9 columns

* 768 Observations
* 9 Attributes

**Features:**

* **Predictors:**
  + Pregnancies - Number of times pregnant
  + Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  + Blood Pressure - Diastolic blood pressure measured in (mm Hg)
  + Skin Thickness - Measured on the triceps skin fold thickness (mm)
  + Insulin - 2-Hour serum insulin measured in (mu U/ml)
  + BMI - Body mass index measured in (weight in kg /(height in m)^2)
  + Diabetes Pedigree Function - Hereditary measurement of diabetes
  + Age - In years
* **Target**: Outcome (1 for Diabetes, 0 for not Diabetes)

***3. Data Preparation. Detect and correct data quality problems (missing data, noise, outliers, etc.) and transform the data into an appropriate format for data analysis.Describe your data preparation process and report the results obtained.***

There was missing data filled in with 0 found in Blood Pressure, Skin Thickness, BMI, Insulin, and Glucose predictors. For each of the previous: glucose had 5 missing values, blood pressure had 35 missing values, skin thickness had 227 missing values, insulin had 374 missing values and BMI had 11 missing values. Since we only had 768 observations, the number of missing values for insulin and skin thickness was probably enough to drop the attribute altogether, but our dataset was way too small to do so. So for future reference it is probably best to take most of the predictions regarding skin thickness and insulin with a grain of salt. We didn’t want to delete the columns and rows with missing values because the data wasn’t very large and we didn’t want to risk having little data to work with. Instead, we took the mean of each column with missing data and filled in the missing values with it.

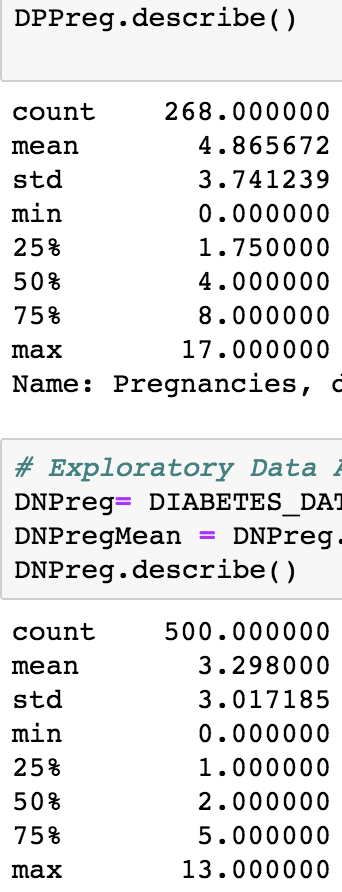
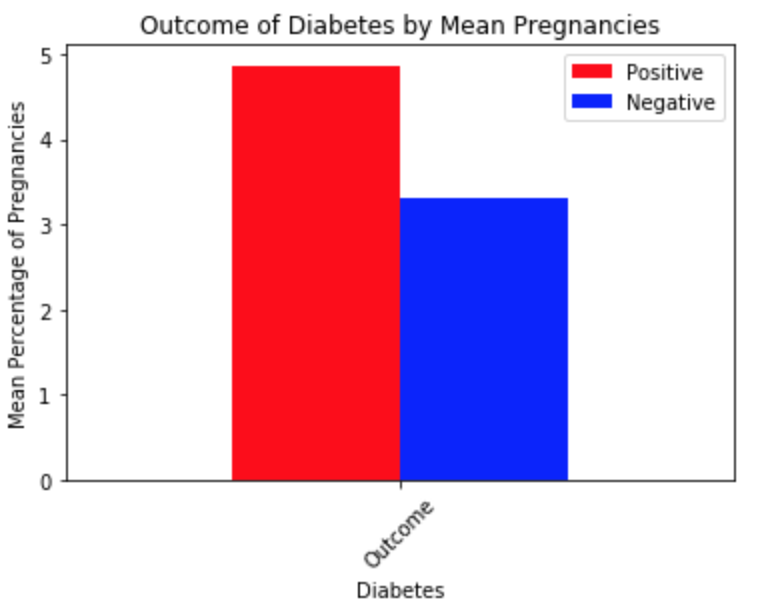
The rest of the data was very clean as there weren’t noticeable outliers or noise.

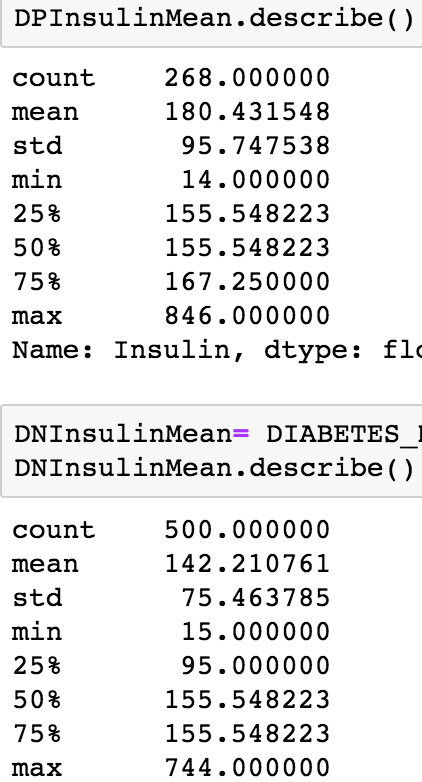
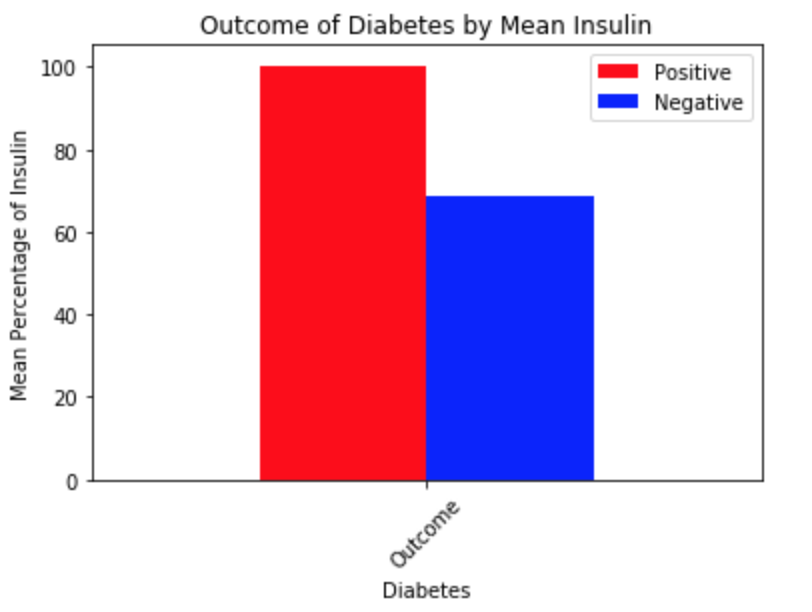
***4. Exploratory Data Analysis. Explore the data using summary statistics and plots and identify the most important variables for data analysis. Describe your exploratory data analysis process and report the results obtained.***

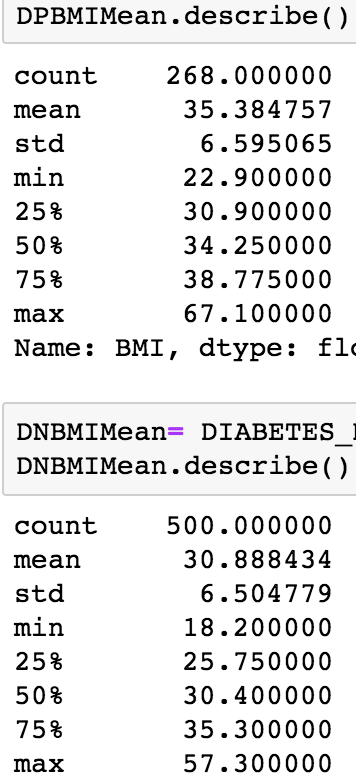
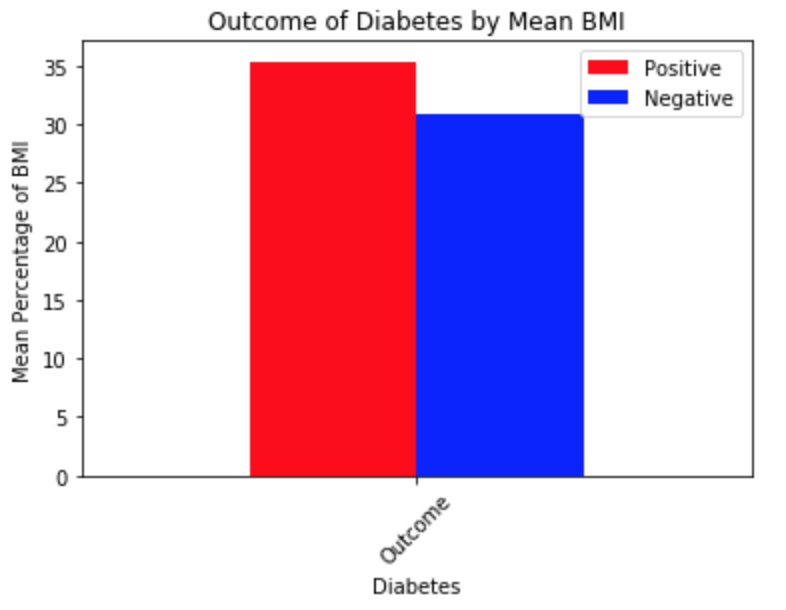
Our goal was to look for data attributes that had a significant or noticeable difference in means for those with diabetes (positive) and those without diabetes (negative). Looking at the graphs below we focused on the these following predictors: Pregnancies, Insulin, Glucose, Diabetes Pedigree Function, BMI, and Age.

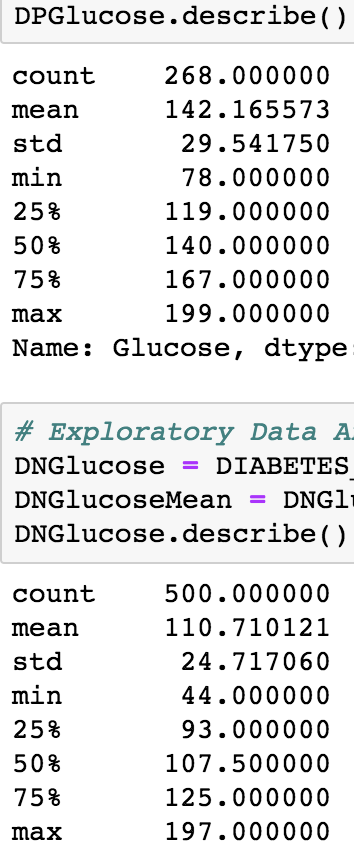
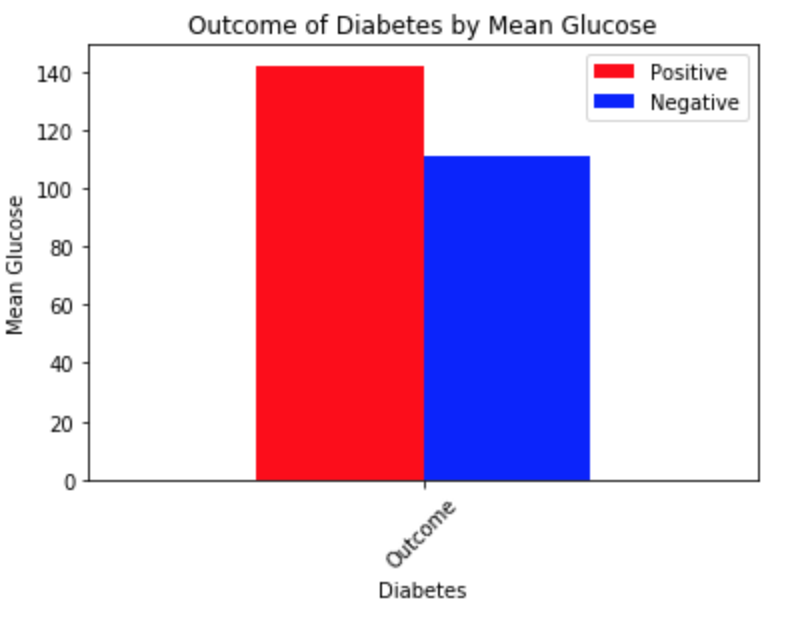
Looking at the charts, in every category those who had diabetes had a higher mean value. This means the people who tested positive for diabetes had more pregnancies, a higher insulin level, a higher glucose level, a high diabetes pedigree function, a higher BMI, and a higher age. We took the biggest differences in the means for this and used them to make predictions in both linear regression and classification.

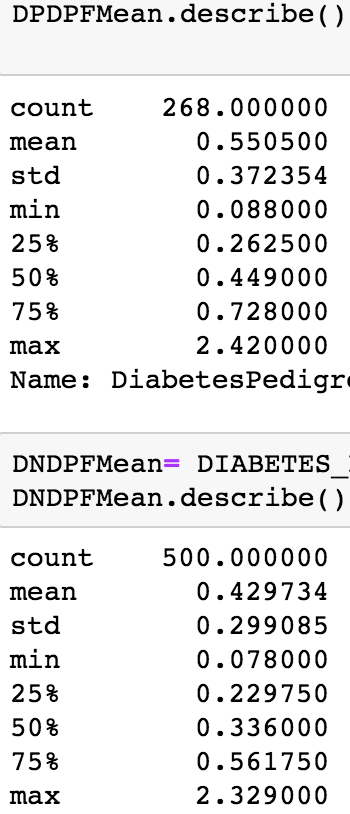
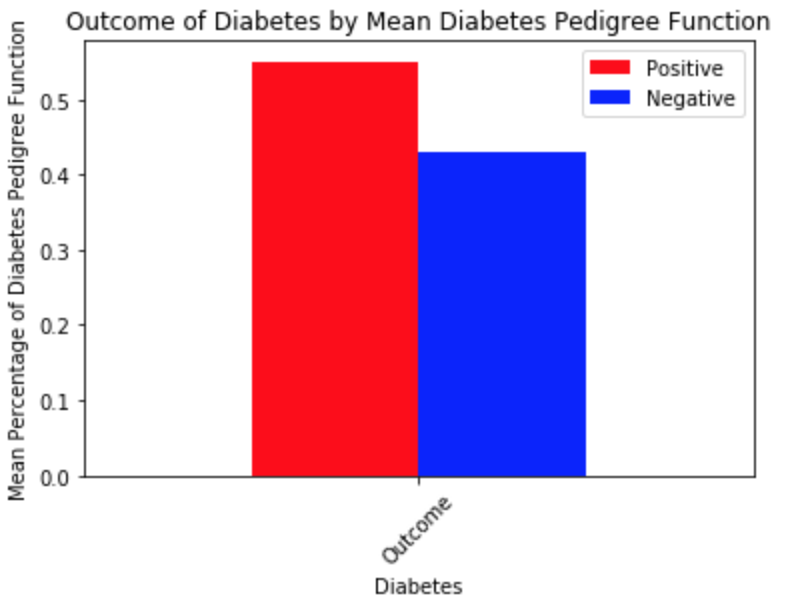
We wanted to make models predicting BMI, glucose, and blood pressure after exploring them and we also used this to find the best variables to predict these. We found age, BMI, glucose, blood pressure to be good variables to explore in regression.

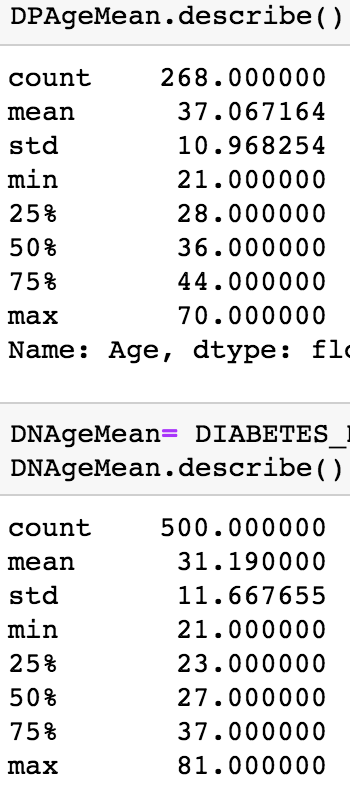
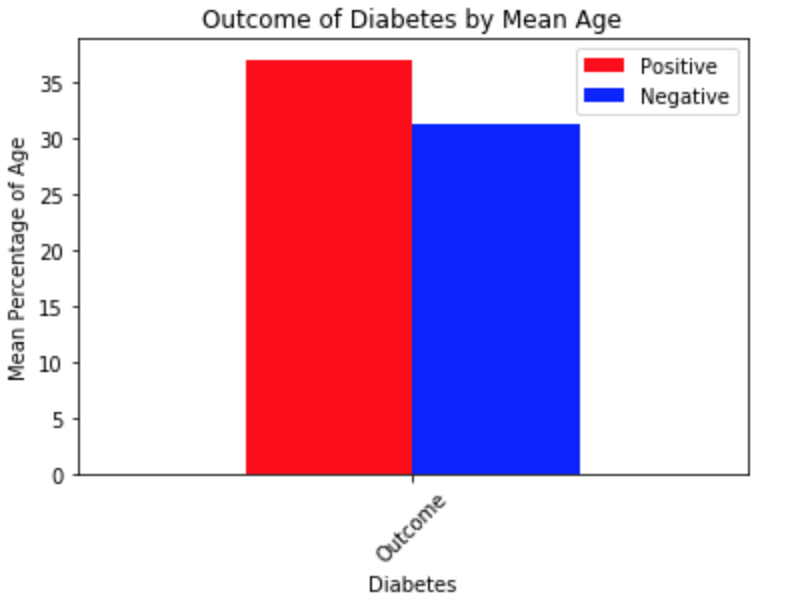
******

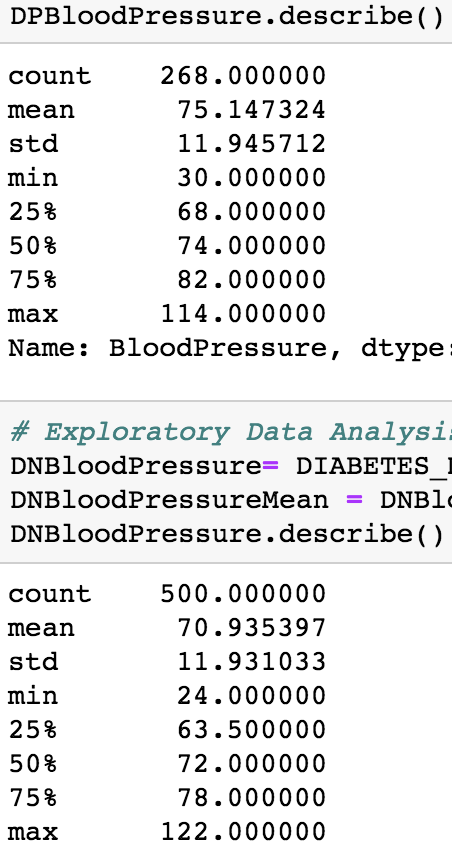
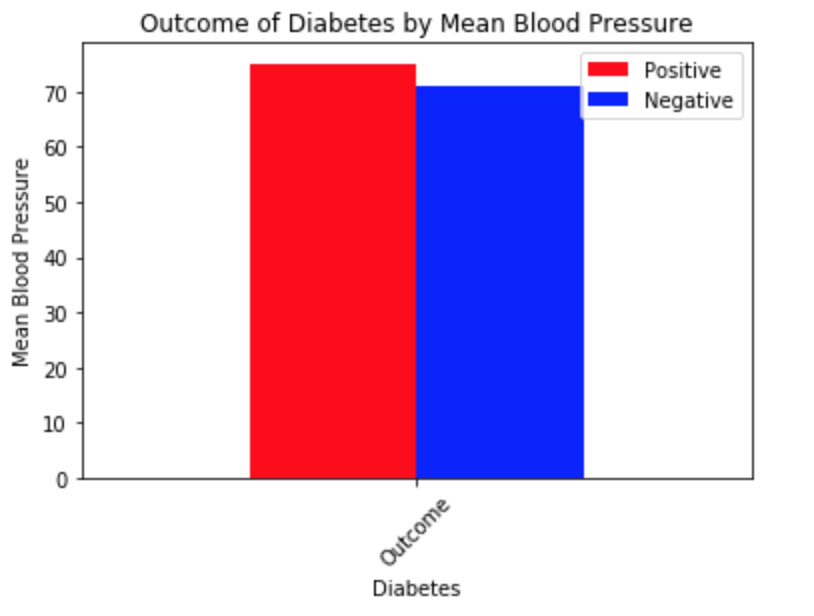
******

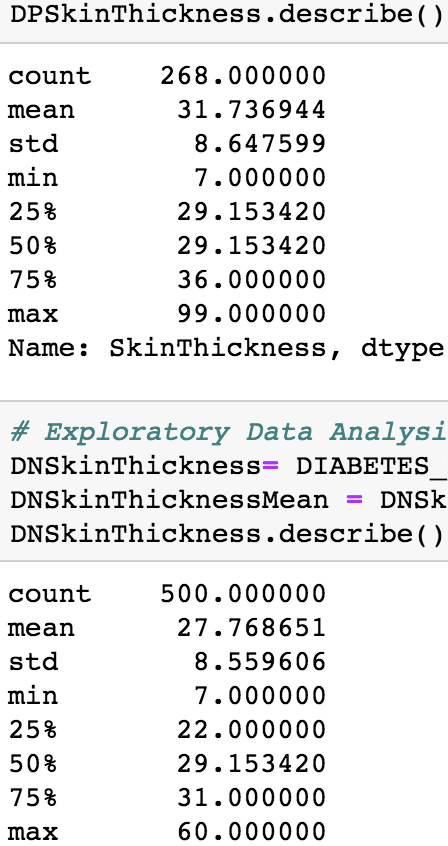
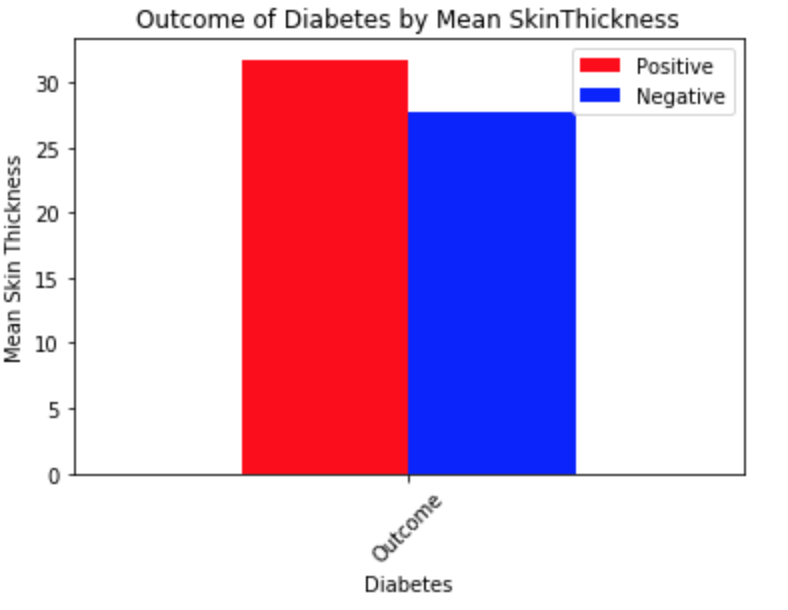
******

******

******

******

******

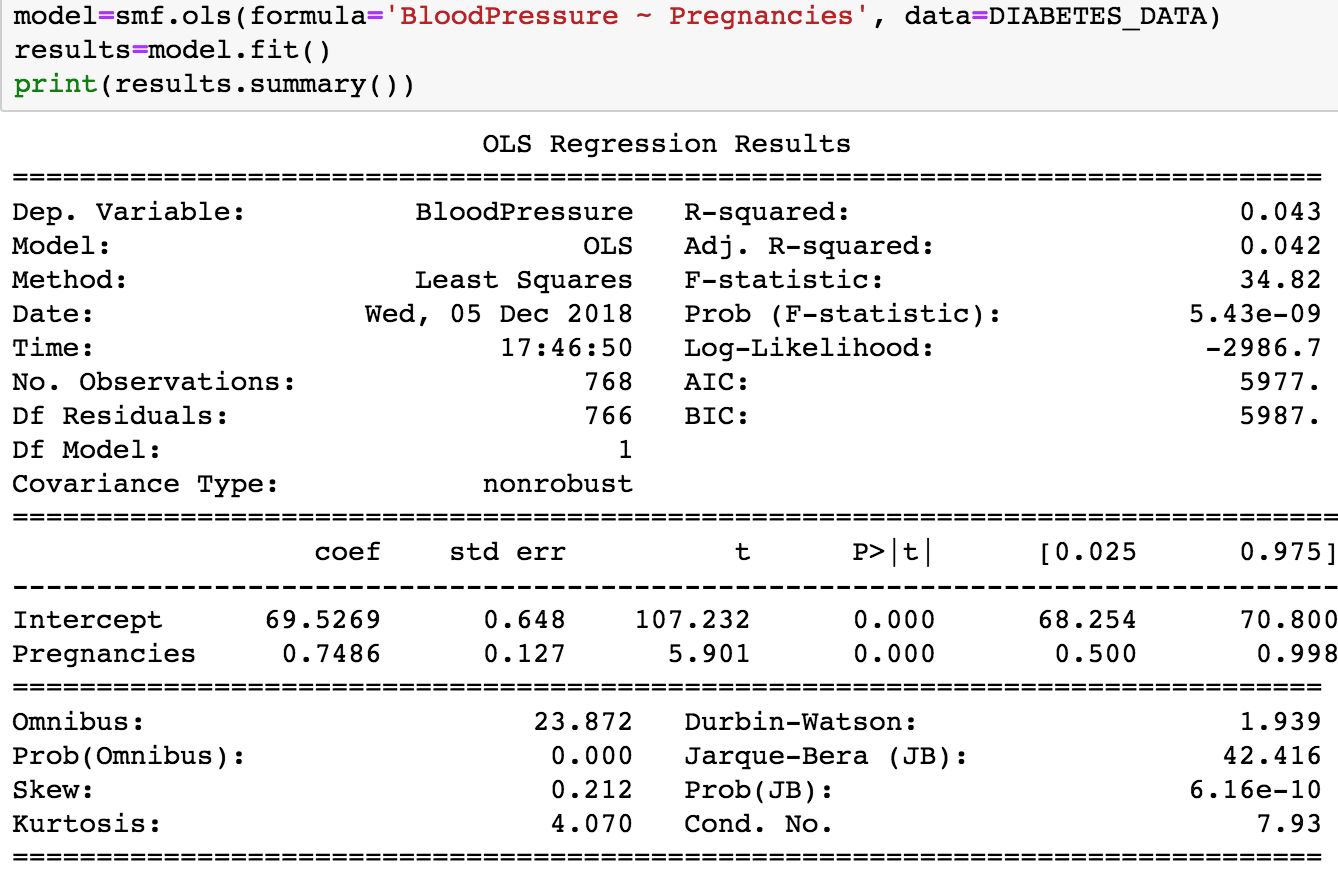
******

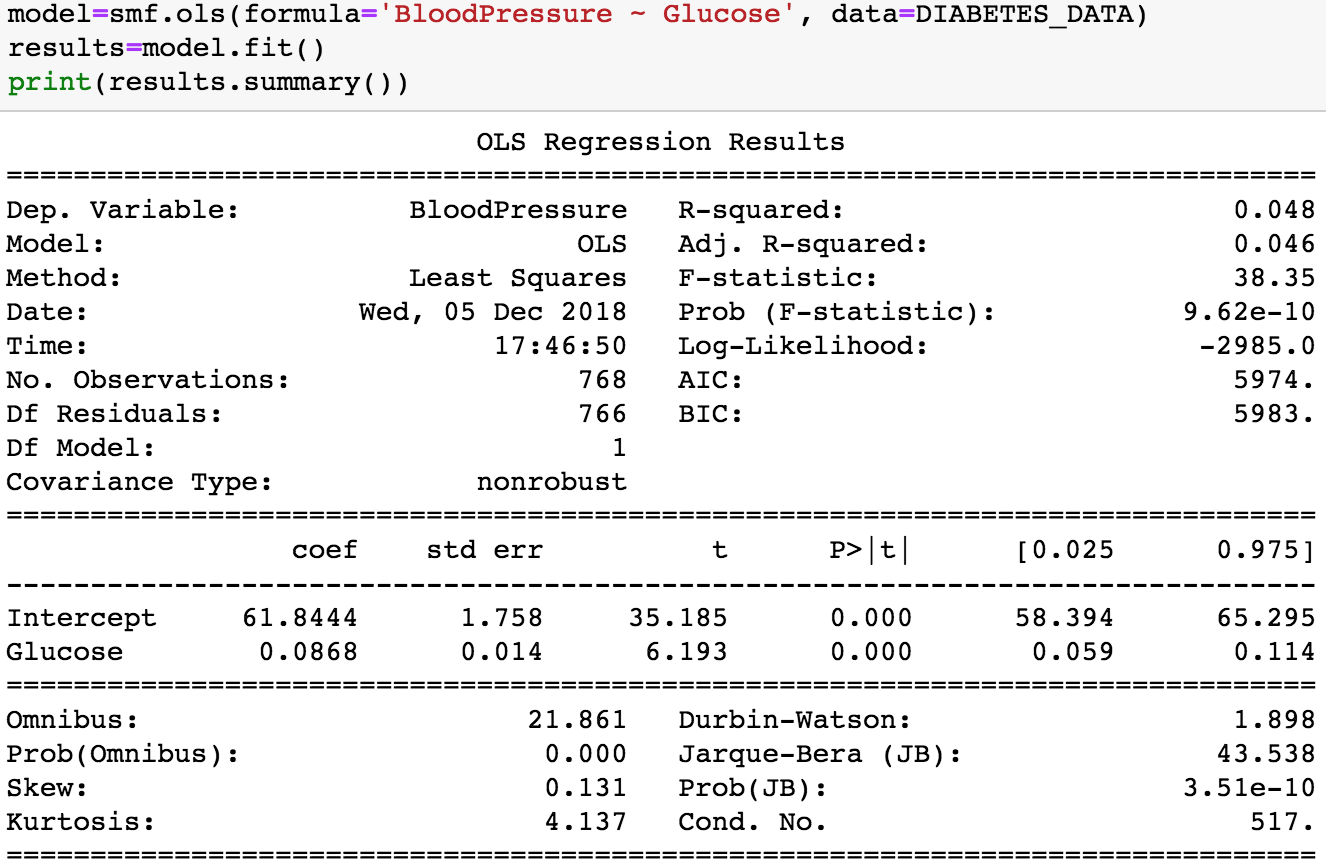
***5. Data Modeling. Train and test one or more models using the data. Your data modeling step must include at least two of the following tasks: (1) Regression, (2) Classification, (3) Clustering, and (4) Text analysis. Consider multiple techniques and multiple parameters. Describe your data modeling process and report the results obtained.***

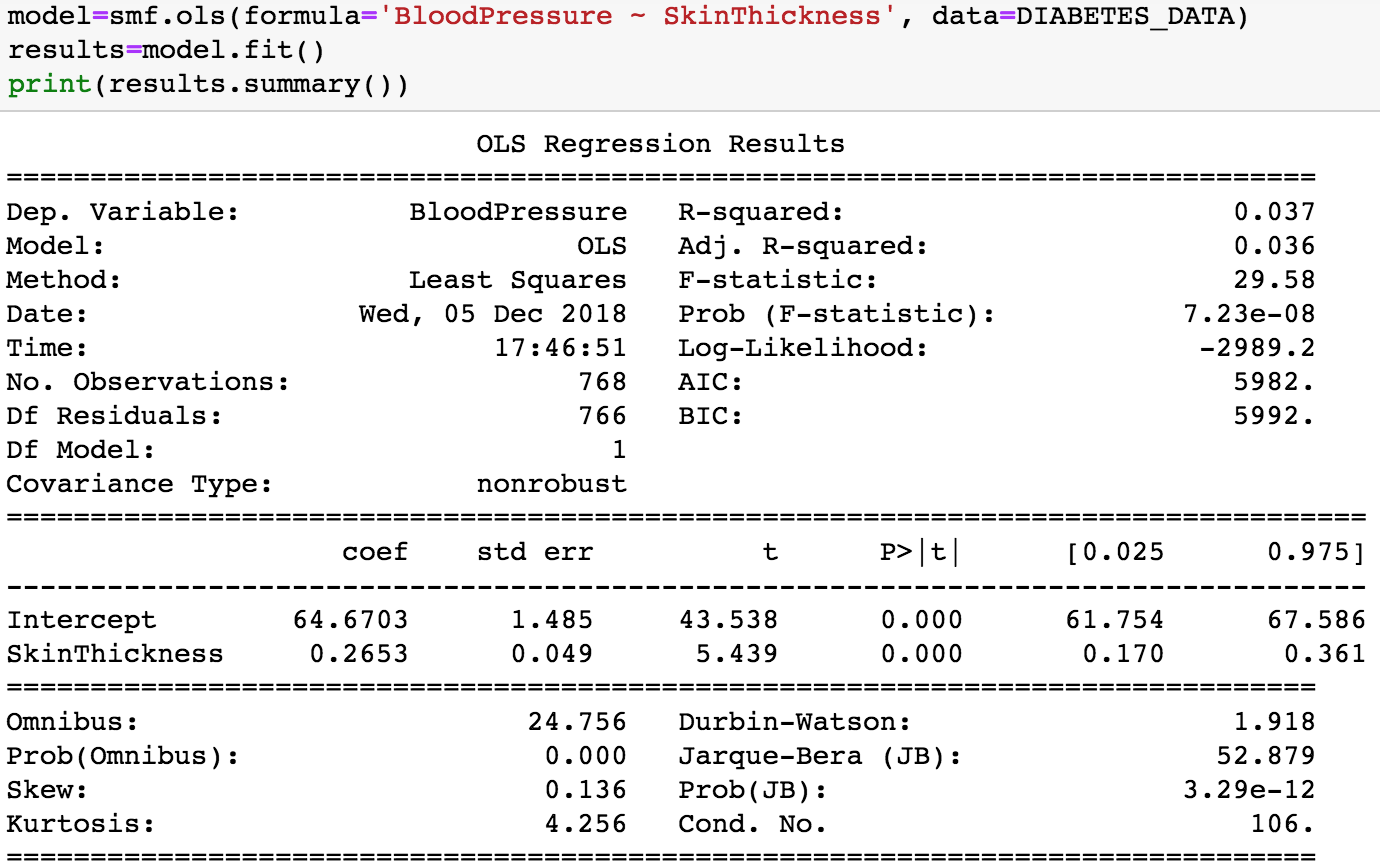
**Regression**

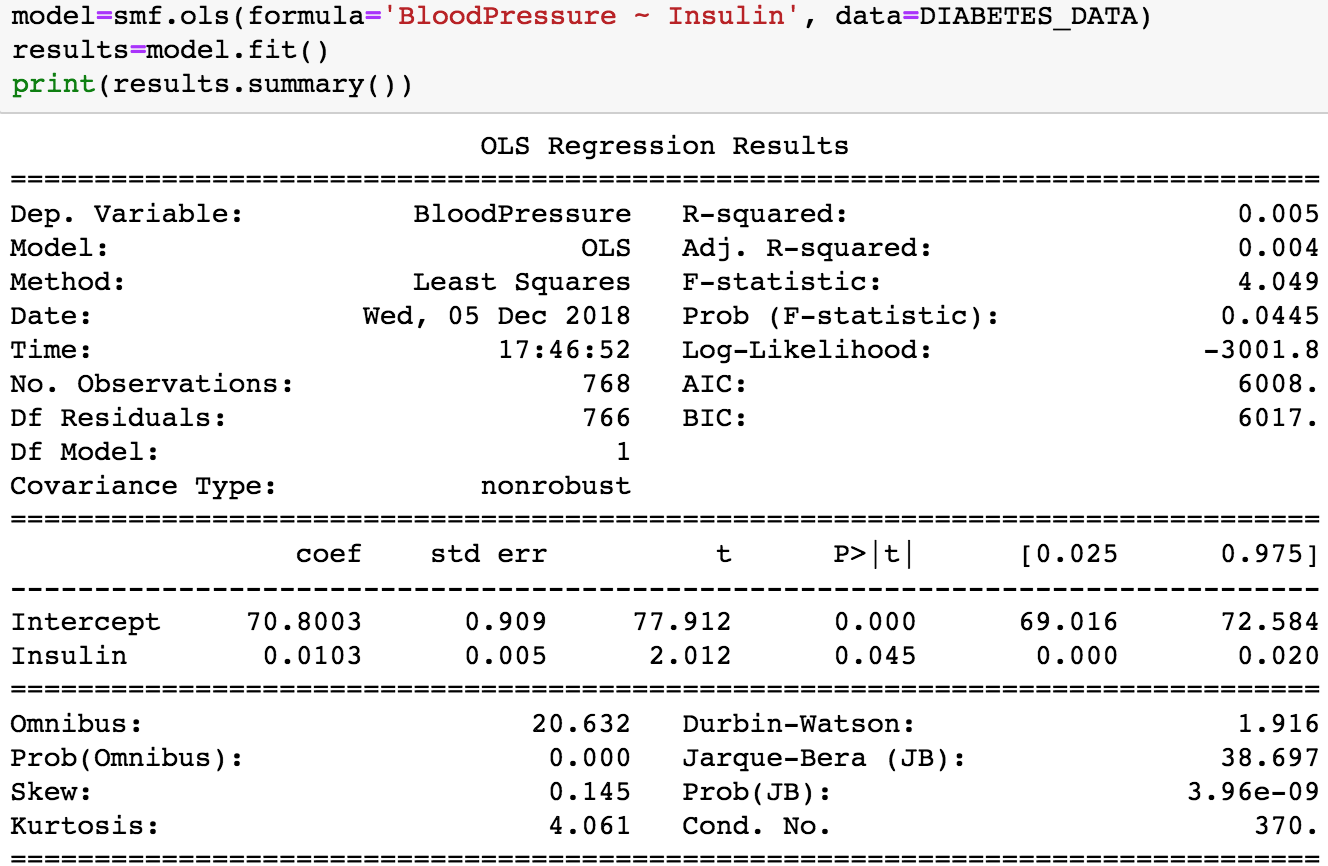
For regression we wanted to test the important quantifiable predictors we thought were contributors to diabetes: blood pressure, glucose, and BMI. We could not predict diabetes in this step so what we wanted to do was see what variables are good at predicting blood pressure, glucose, and BMI, and then using the best predictors of these for a multiple linear regression model used in our models for classification. To do this we tried to predict BMI, glucose, and blood pressure with every other variable in the dataset. We then found the highest R^2 for each which for blood pressure was age .105, BMI .079, and glucose .048, we then did multiple linear regression on this which was .186. For predicting glucose the R^2 was BMI .053, age .071, insulin .177, blood pressure .048 which were put together in multiple linear regression to get an R^2 of .255. For predicting BMI the R^2 of the variables were skin thickness .294, glucose .053, blood pressure .079, which was then combined into multiple linear regression for .336. The highest R^2 value obtained was 0.336, found from the relationship between BMI and Blood Pressure, Skin Thickness, and Glucose. This means that 33% of the variability in BMI can be explained by those 3 variables.

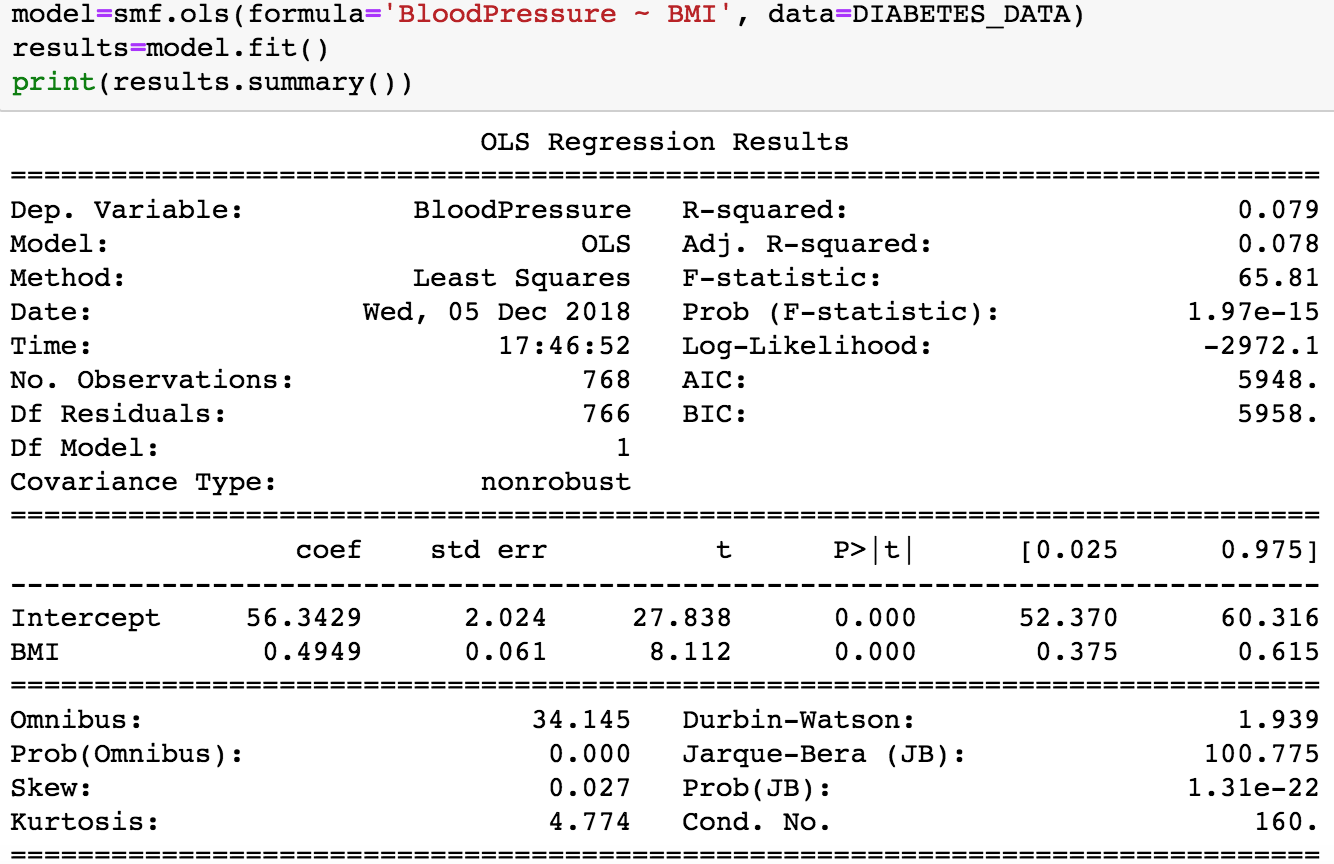
Simple linear regression for blood pressure:

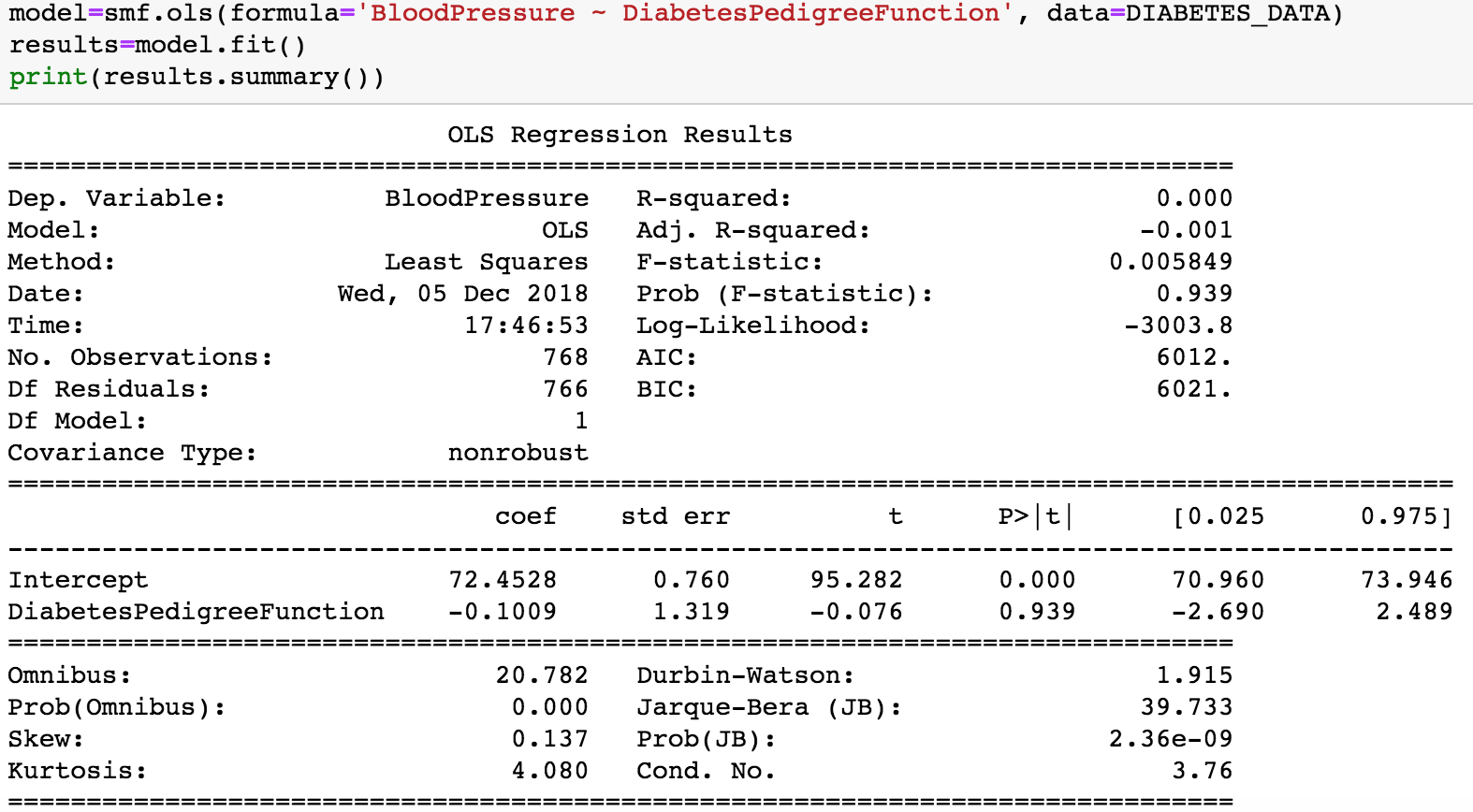


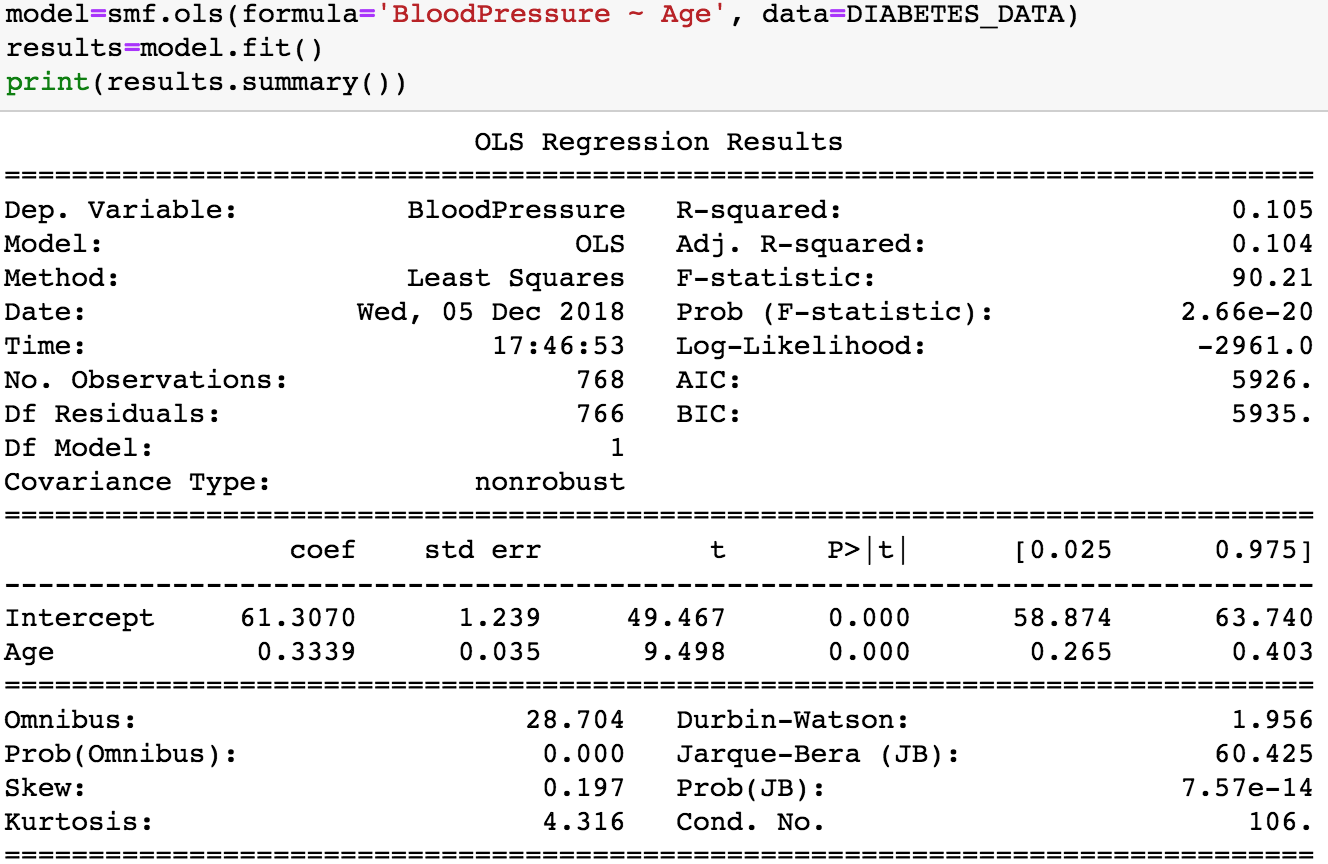




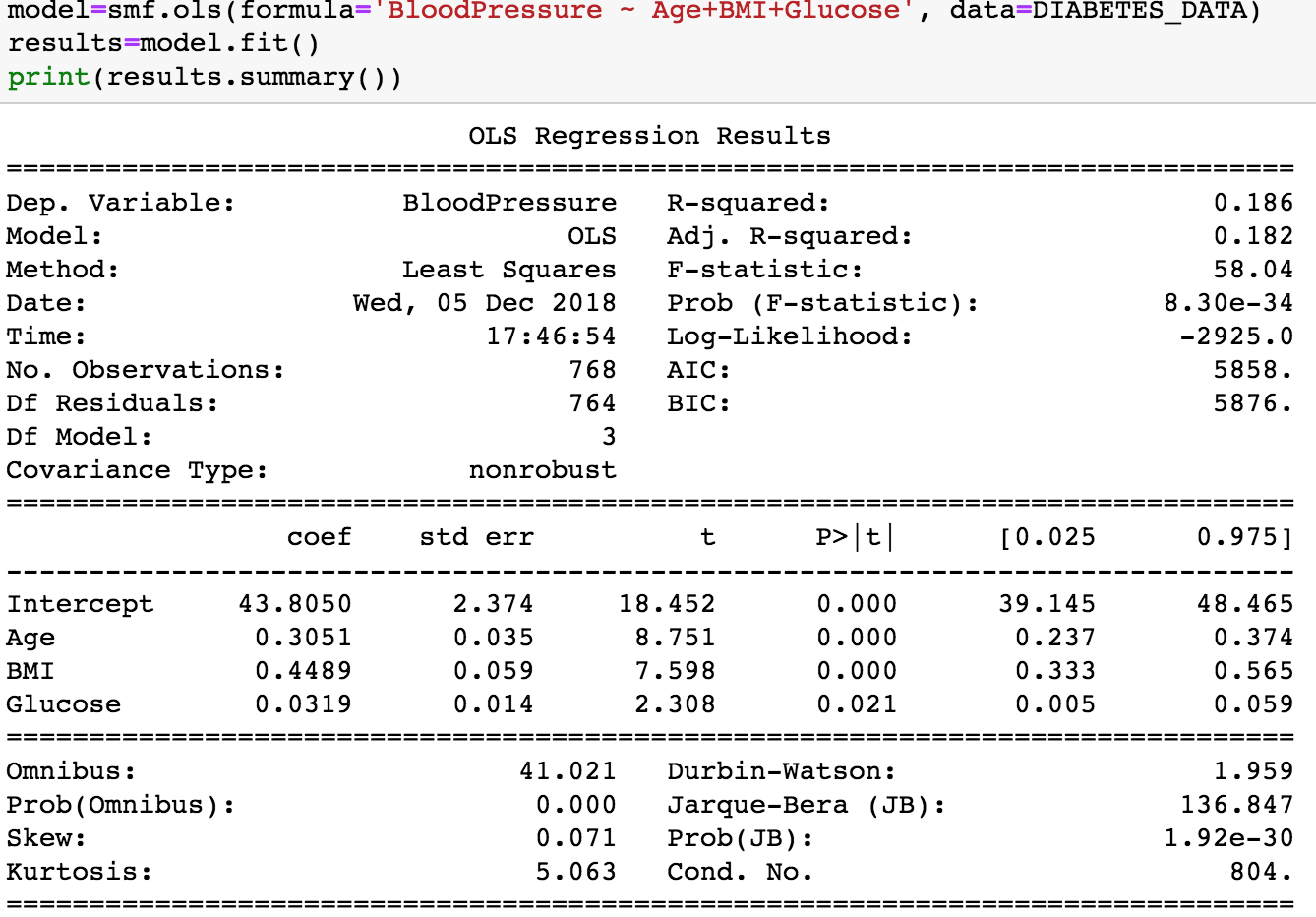




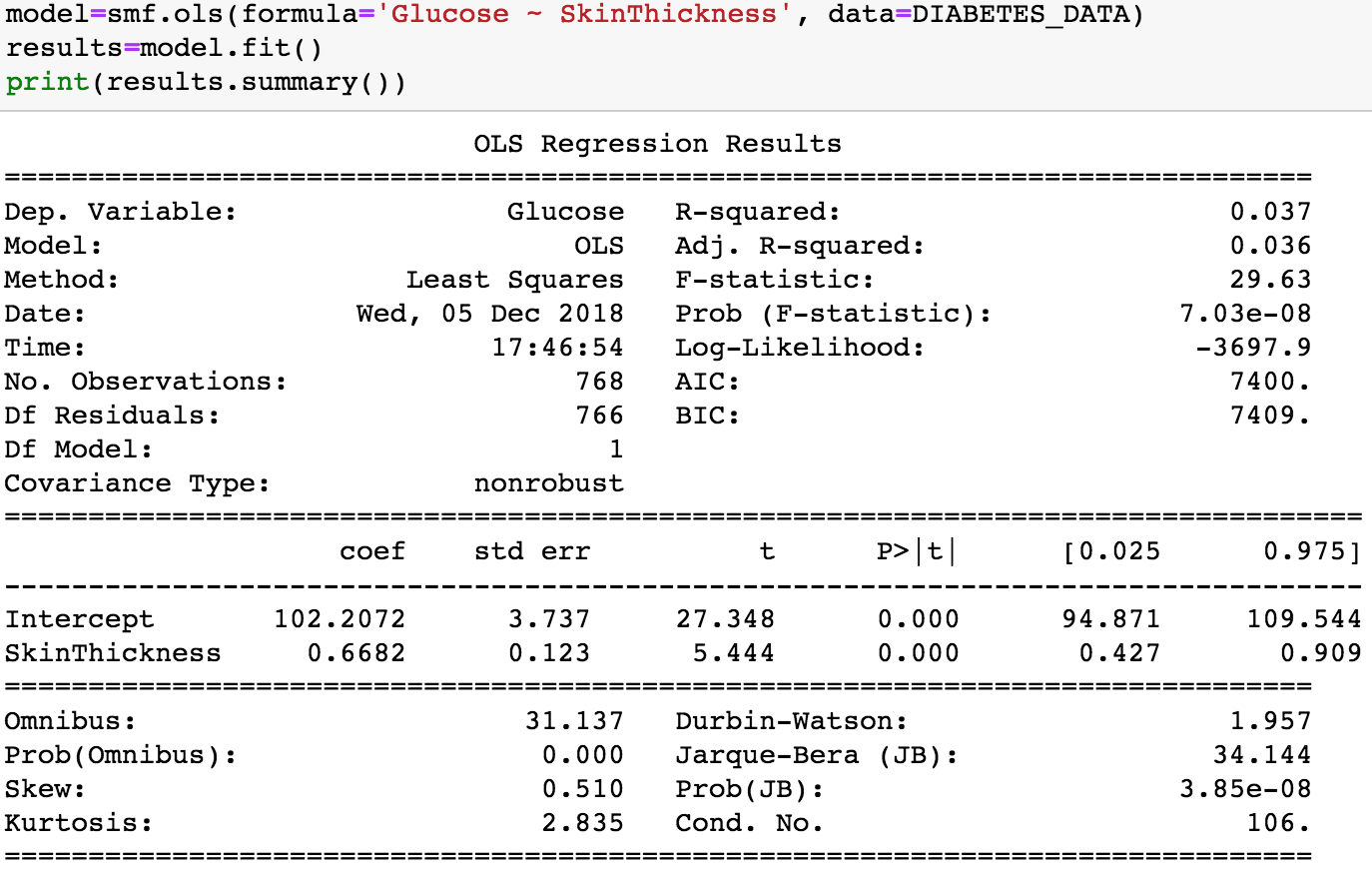




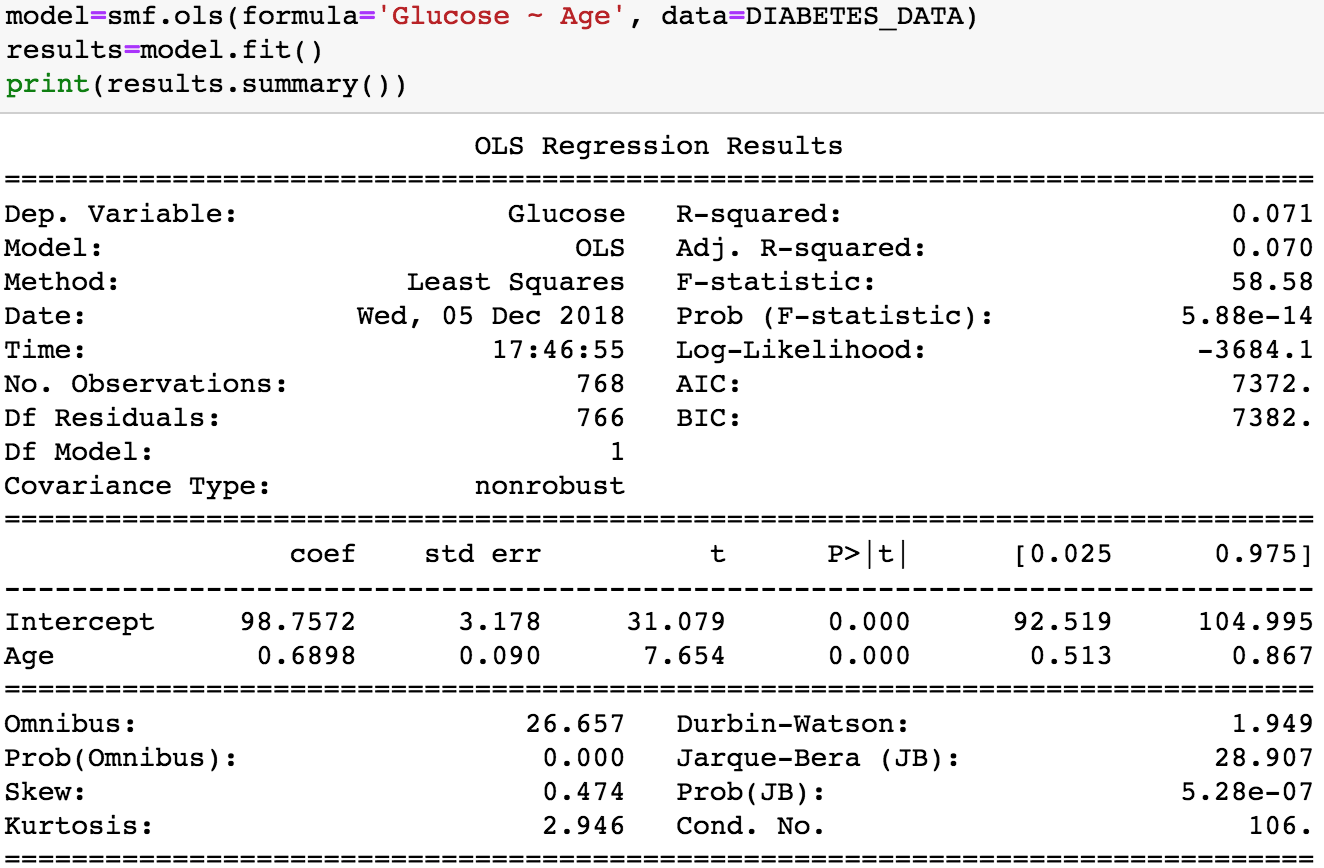
Multiple Linear Regression for Blood Pressure:

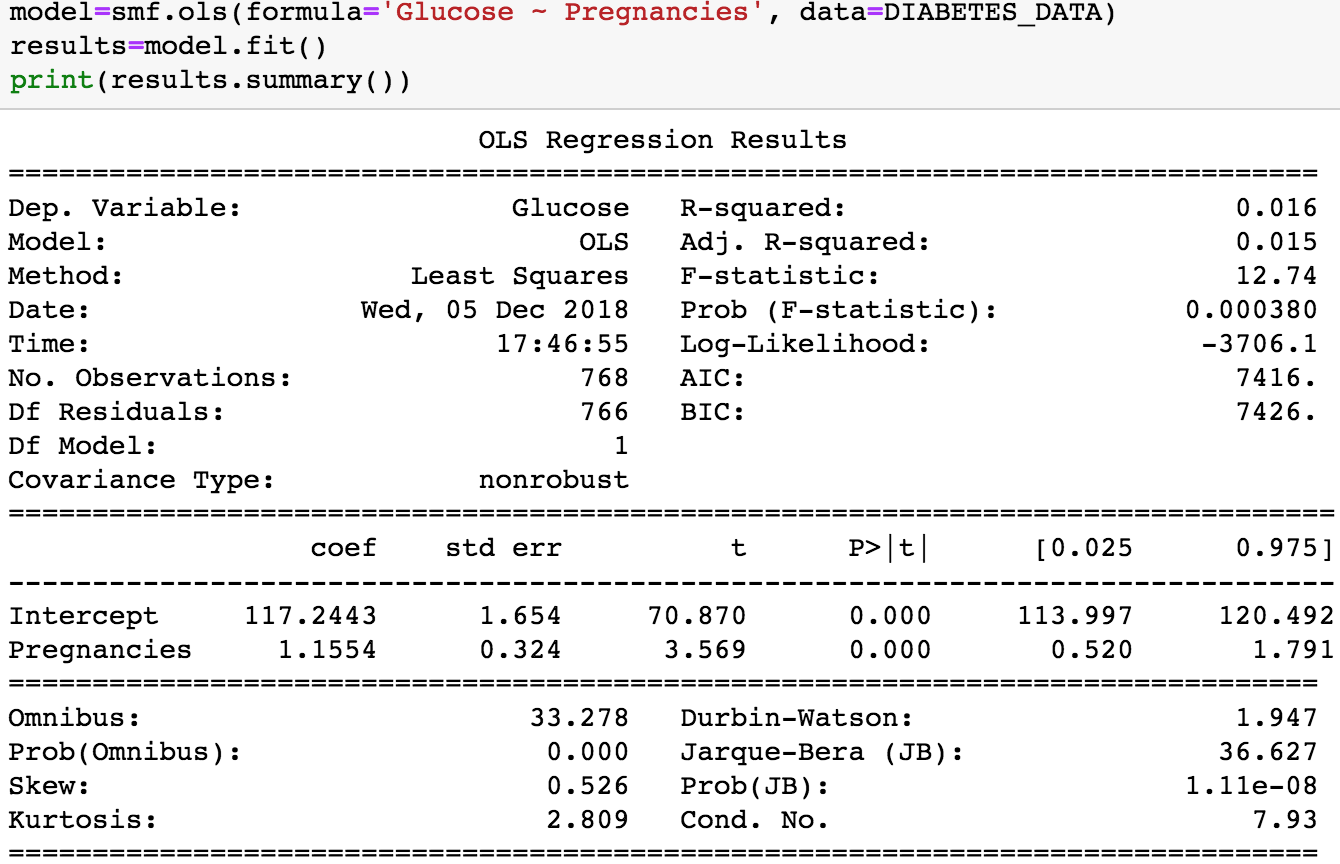


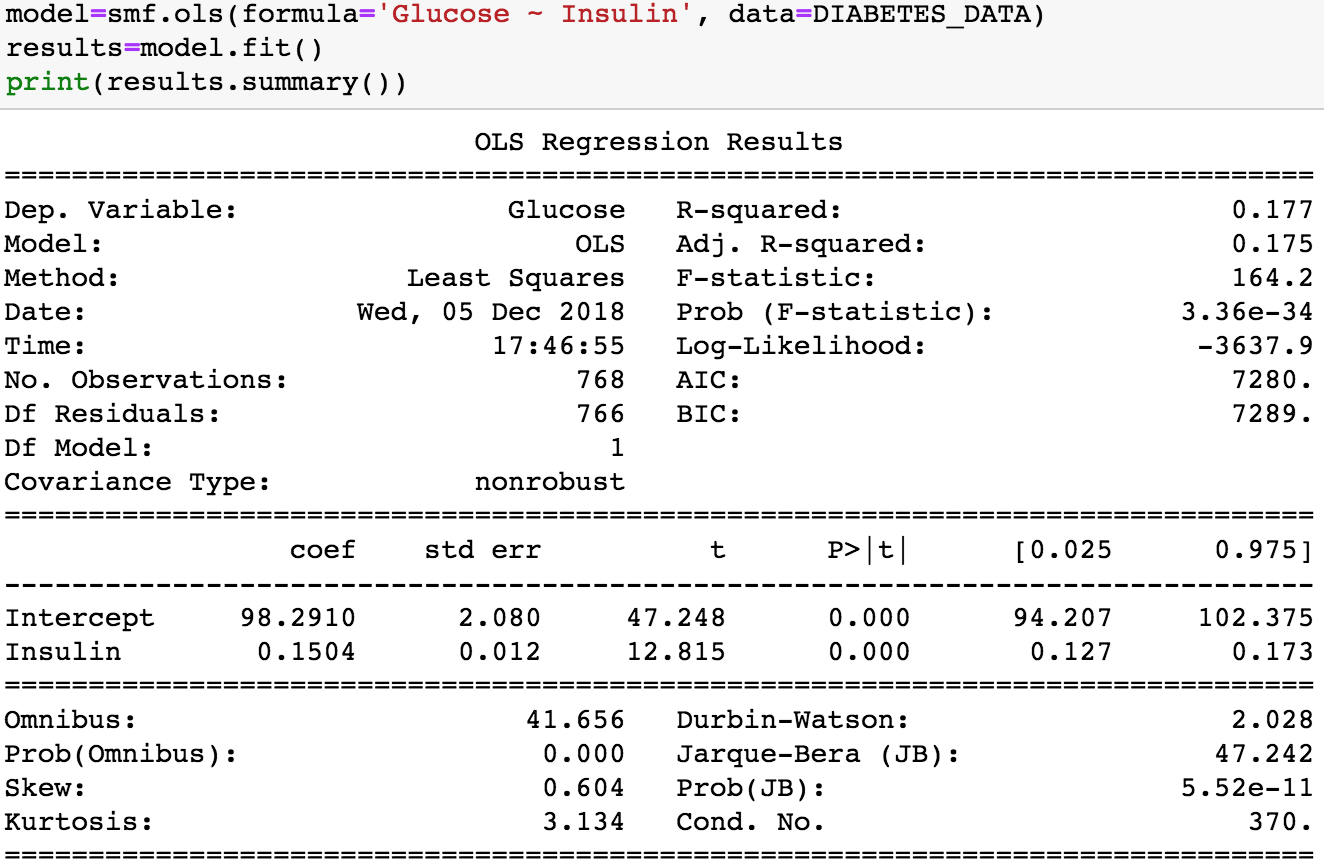
Glucose Single Linear Regression:

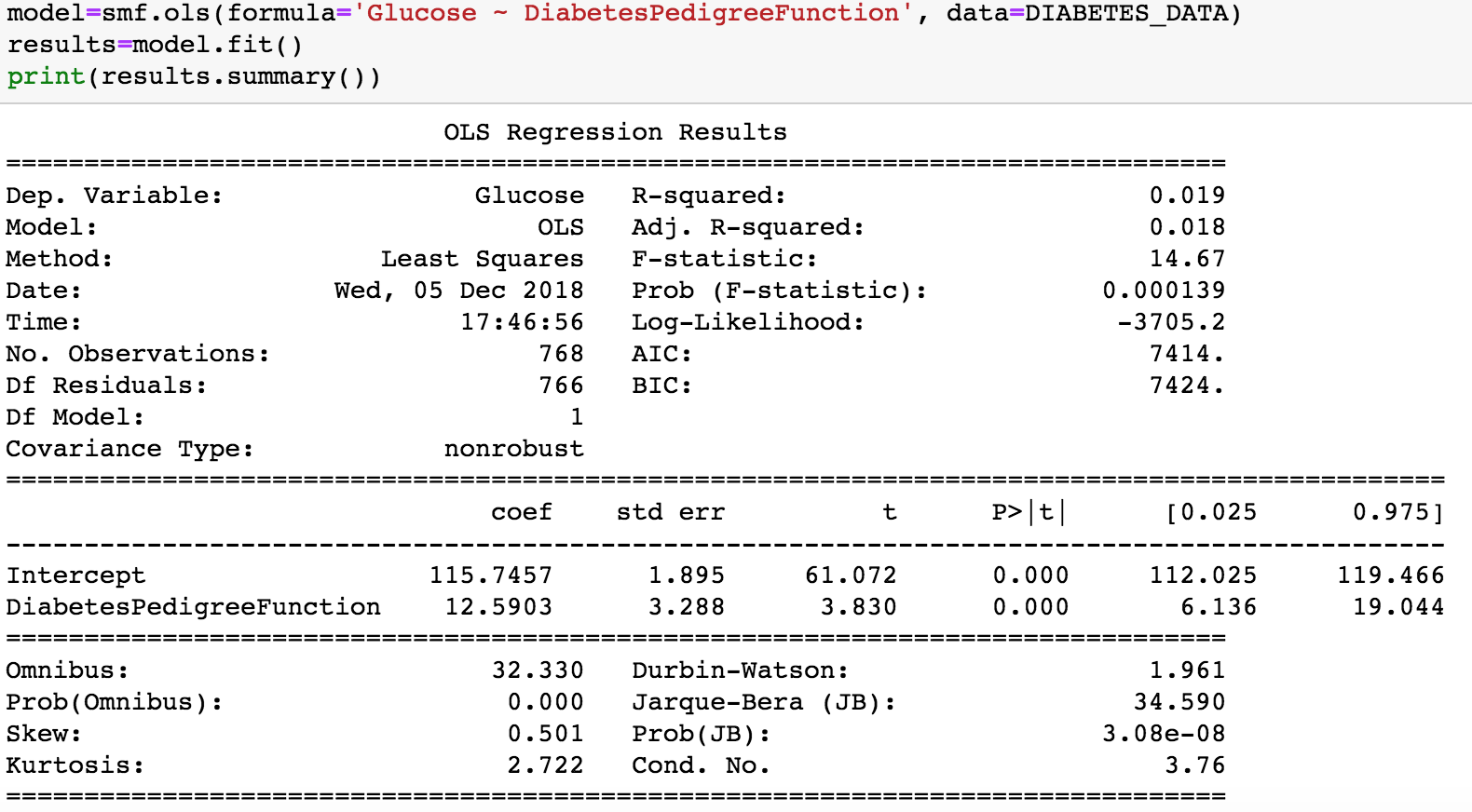


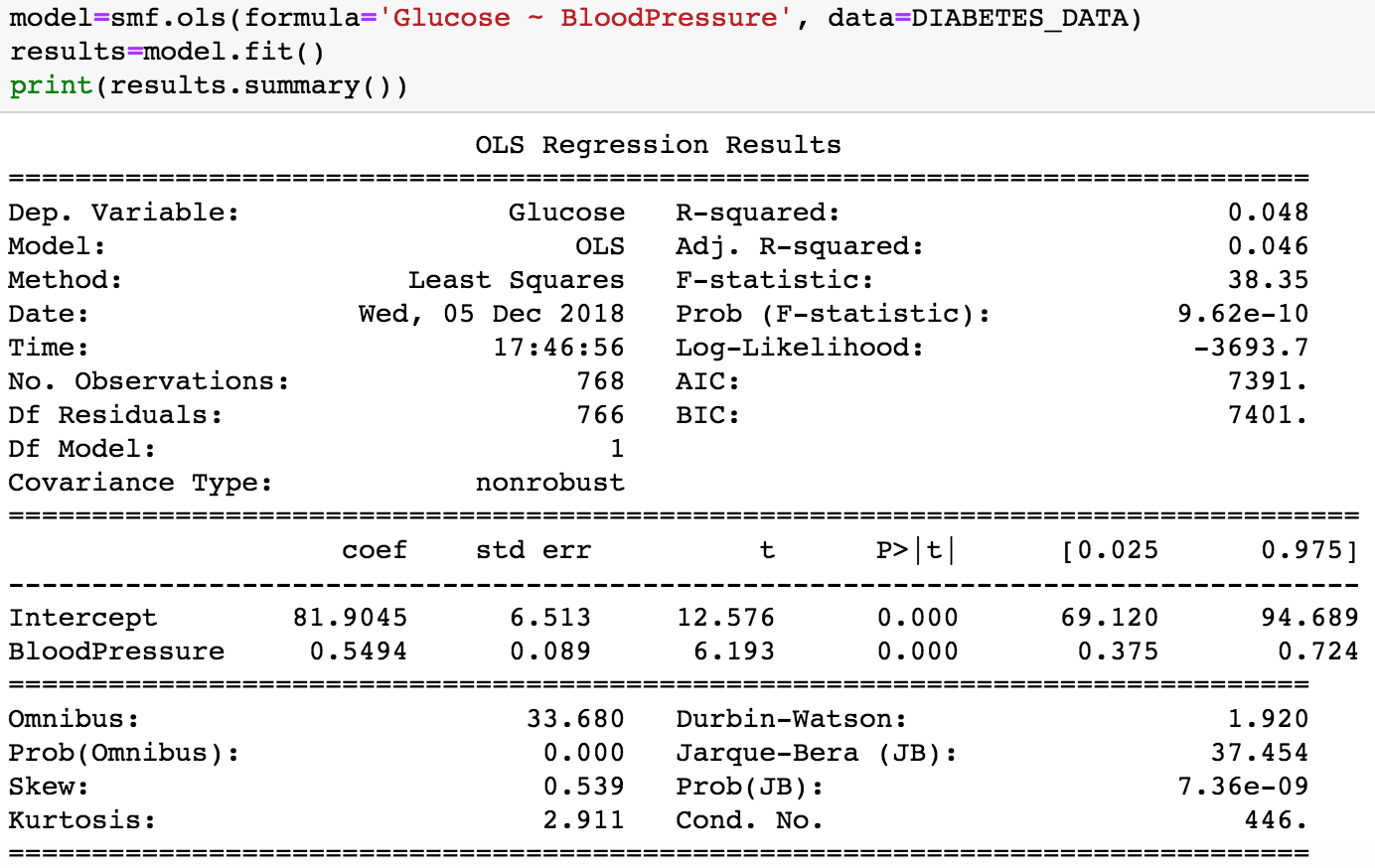
****

****

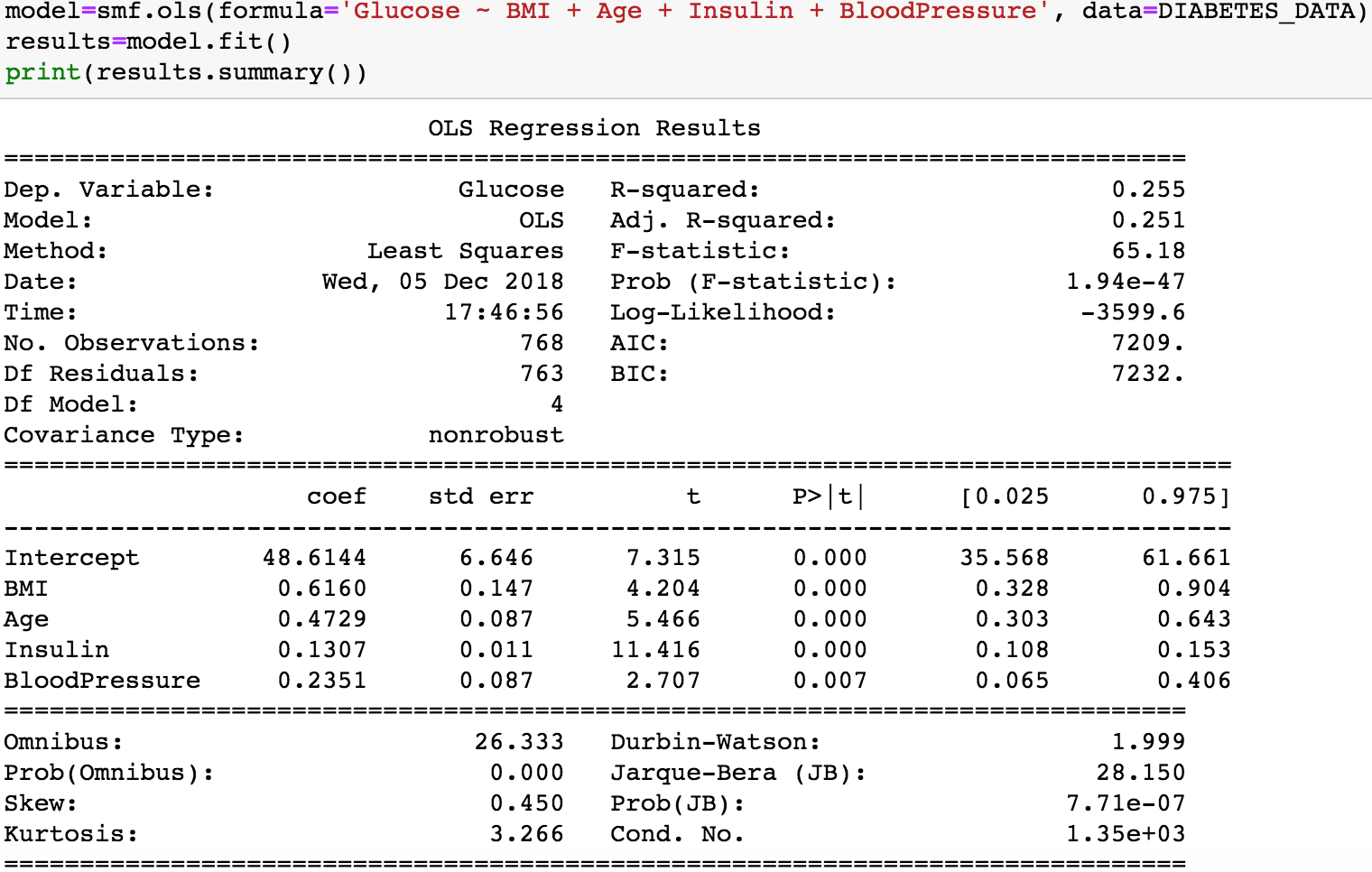
****

****

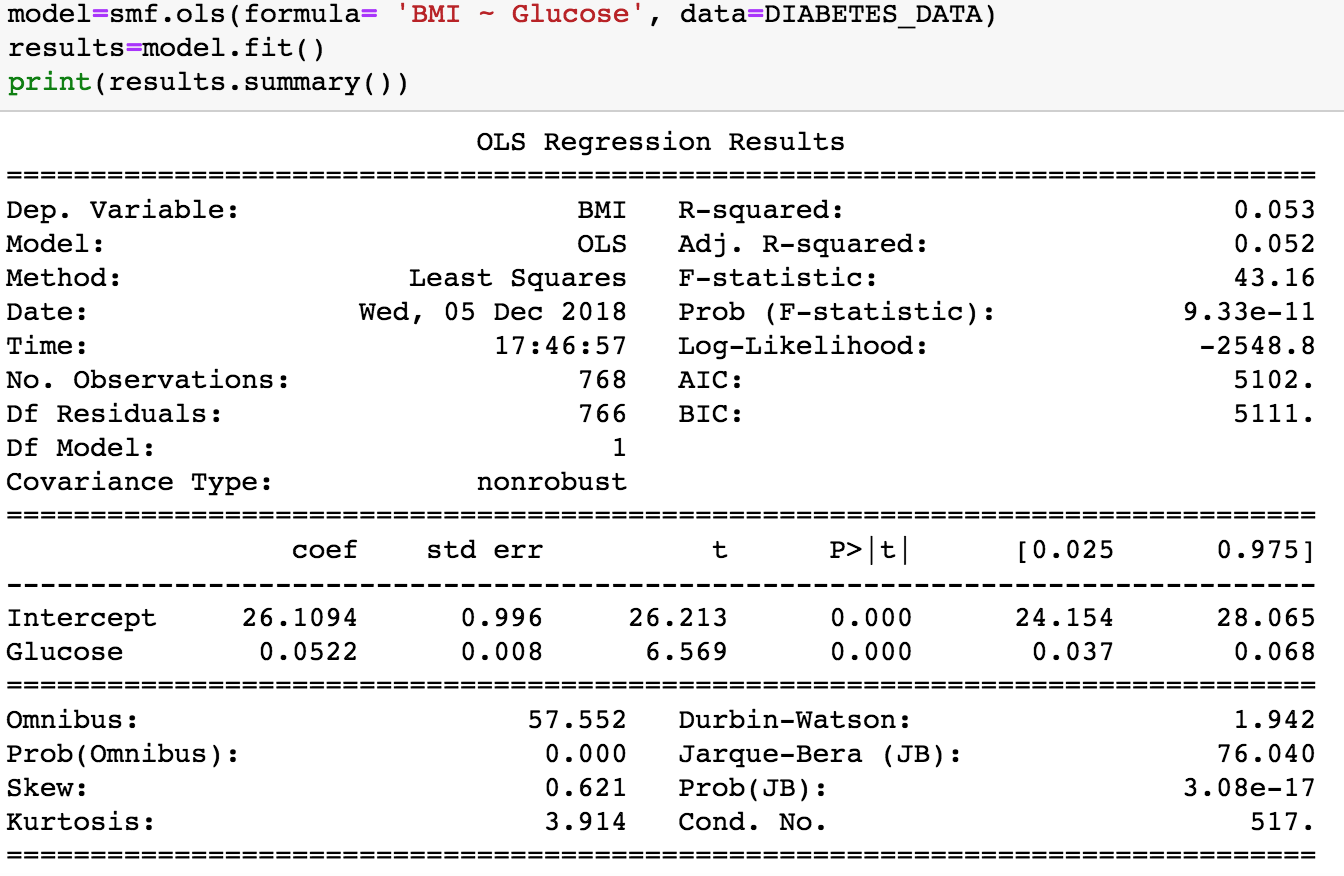
****

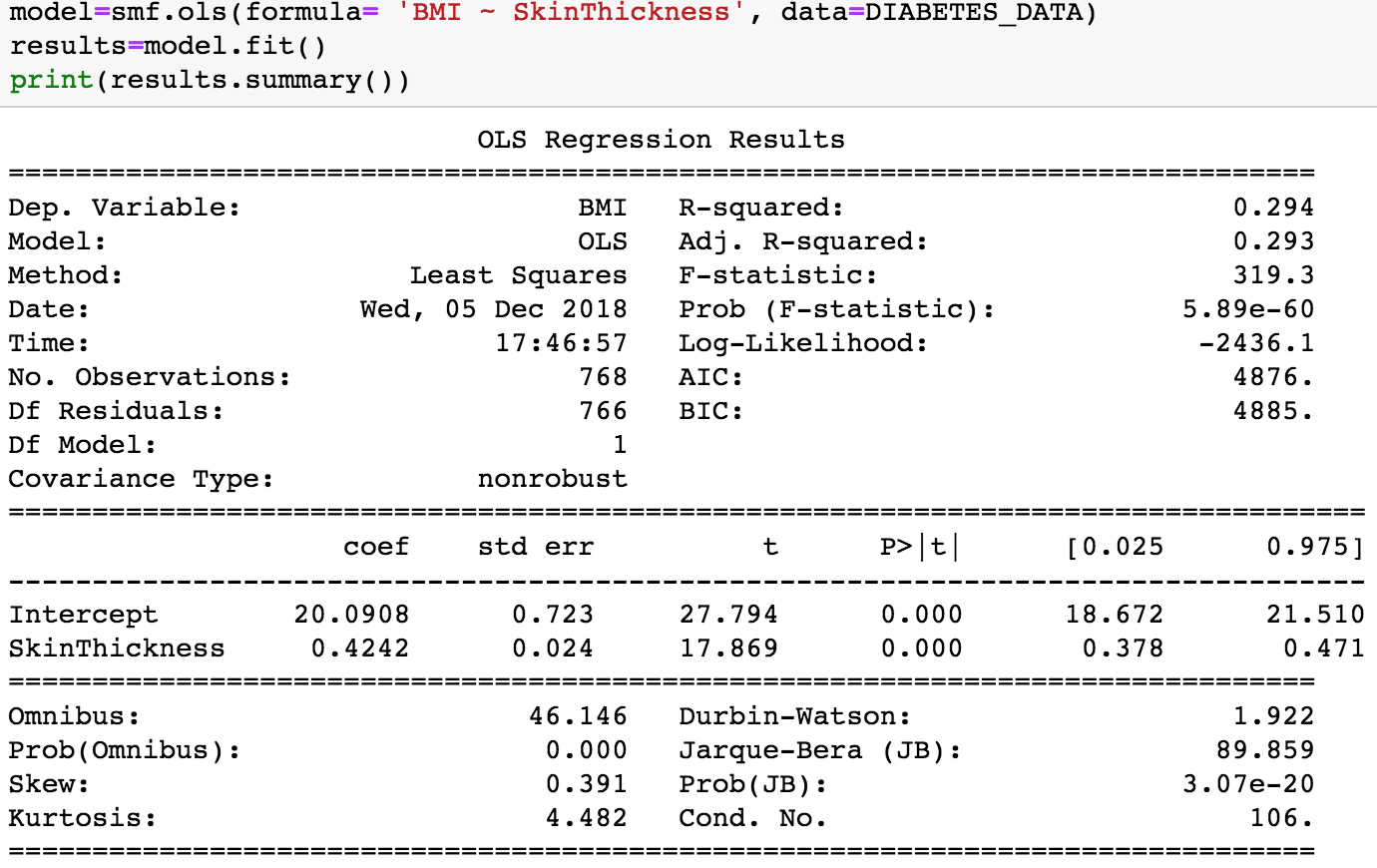
****

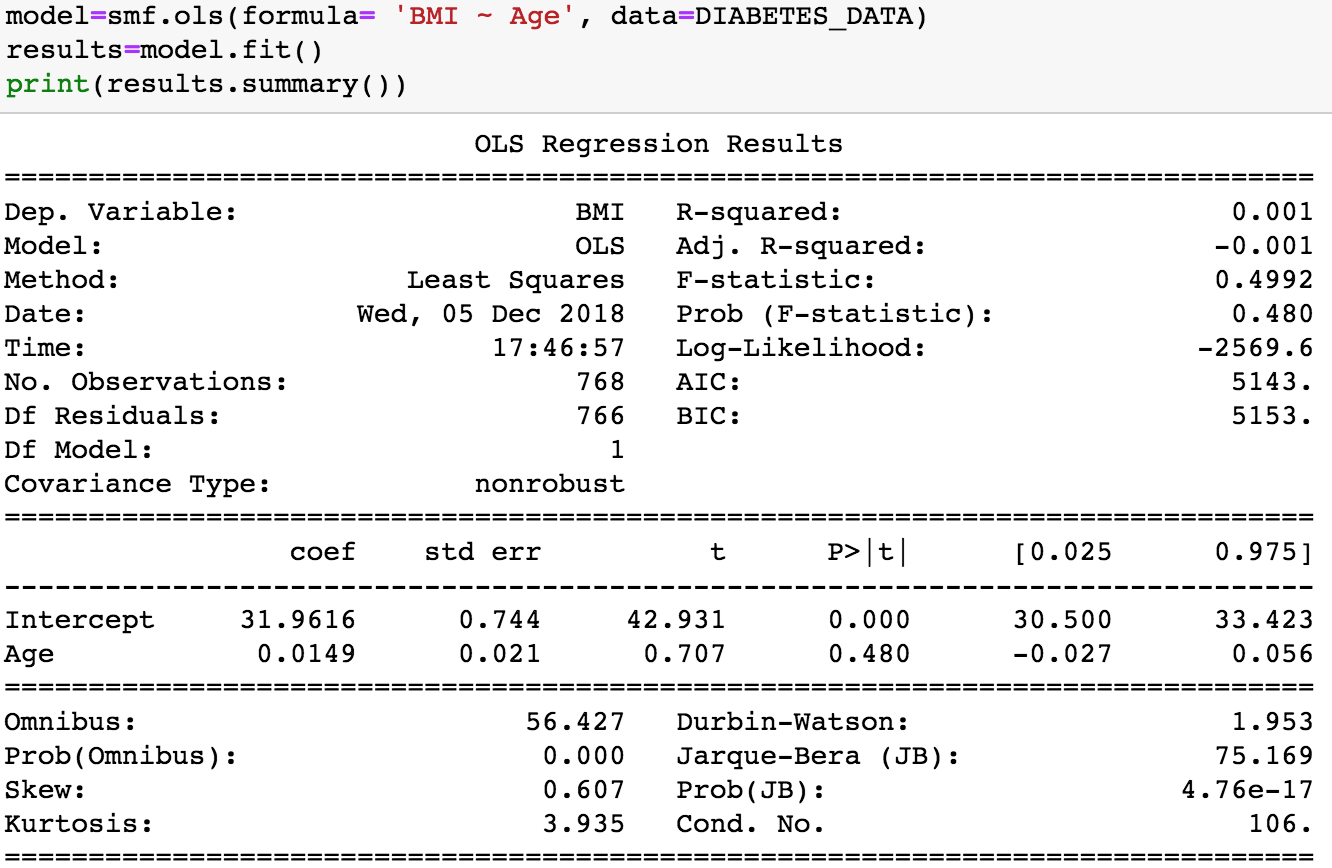
Glucose Multiple Linear Regression:

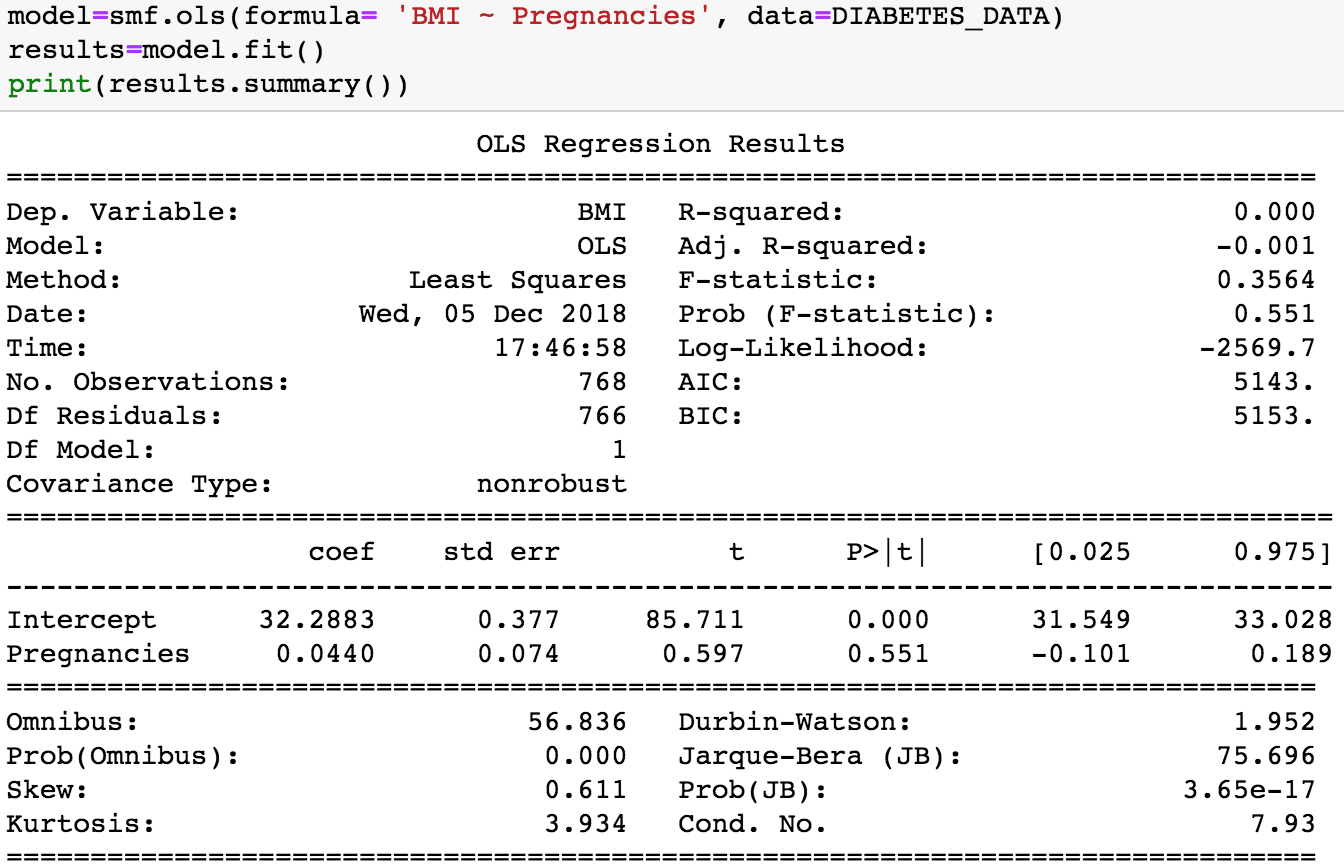


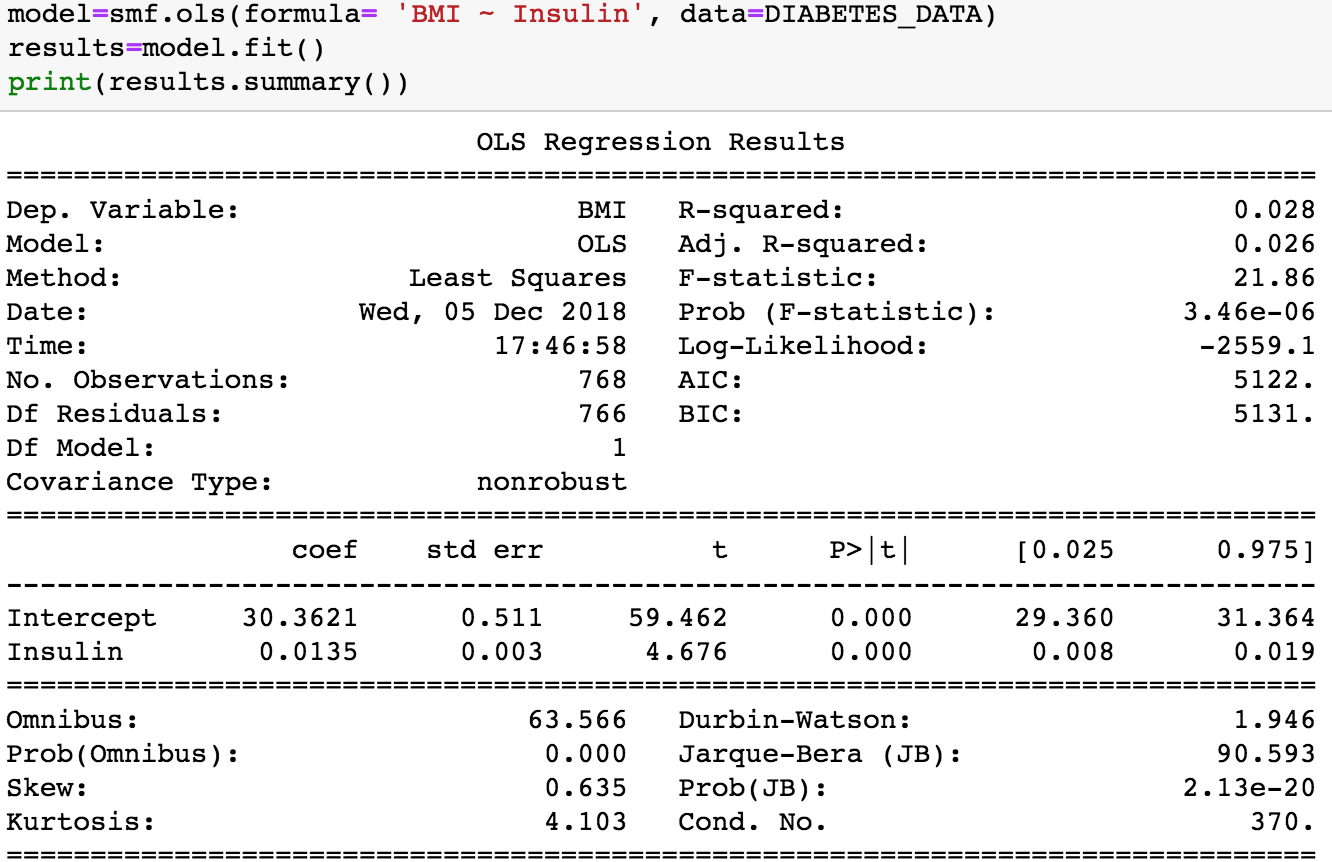
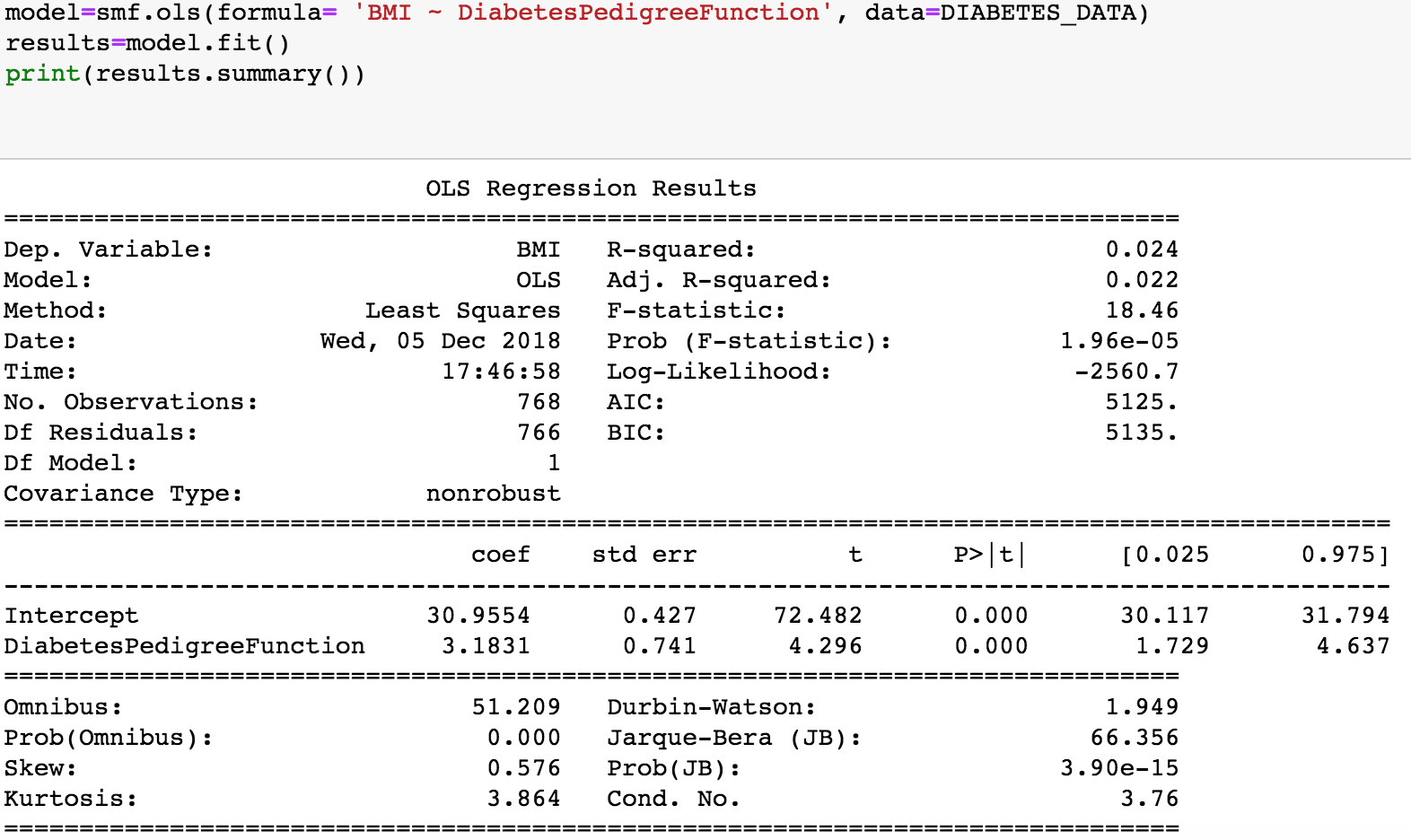
Single Linear Regression BMI:

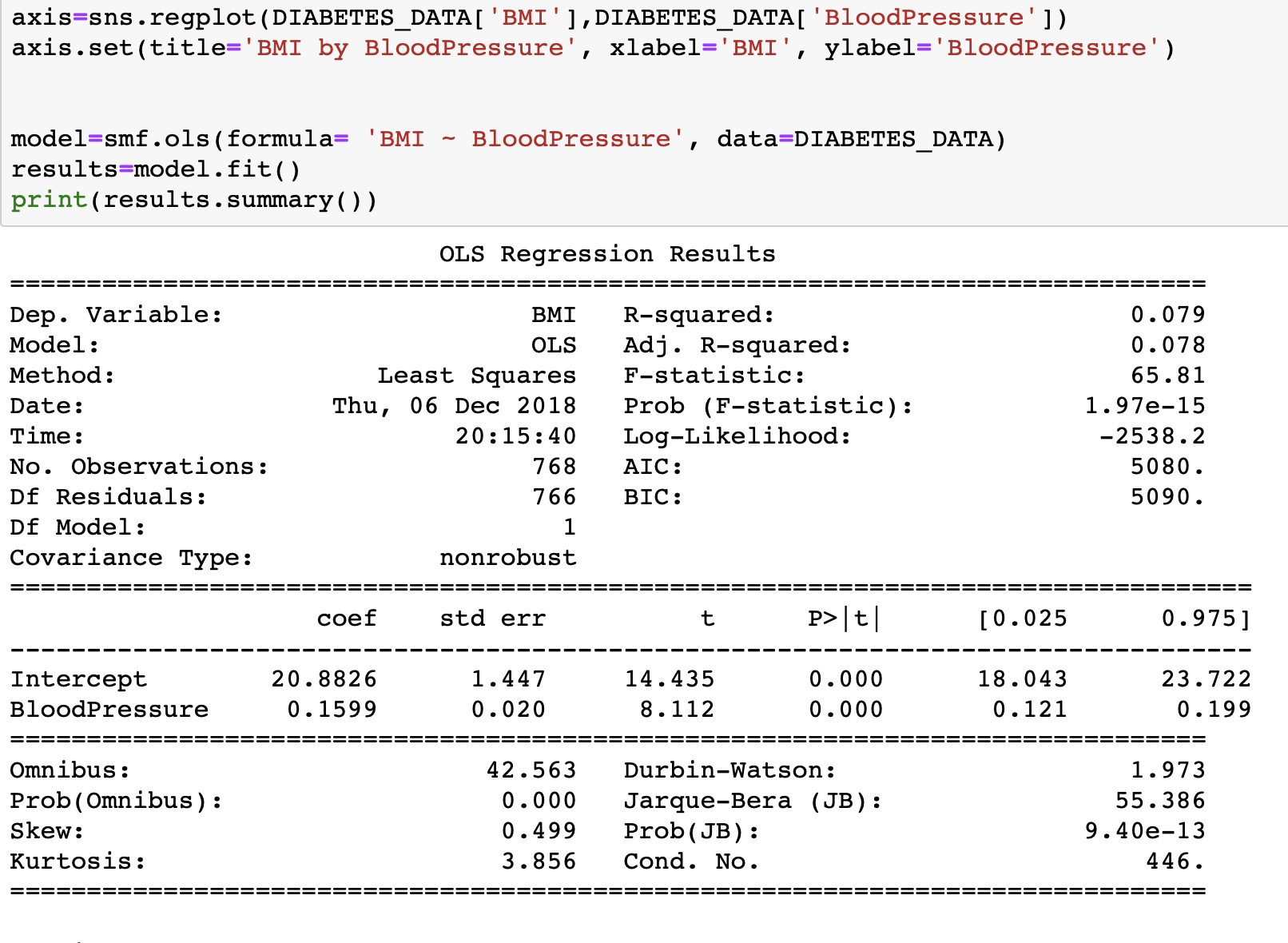




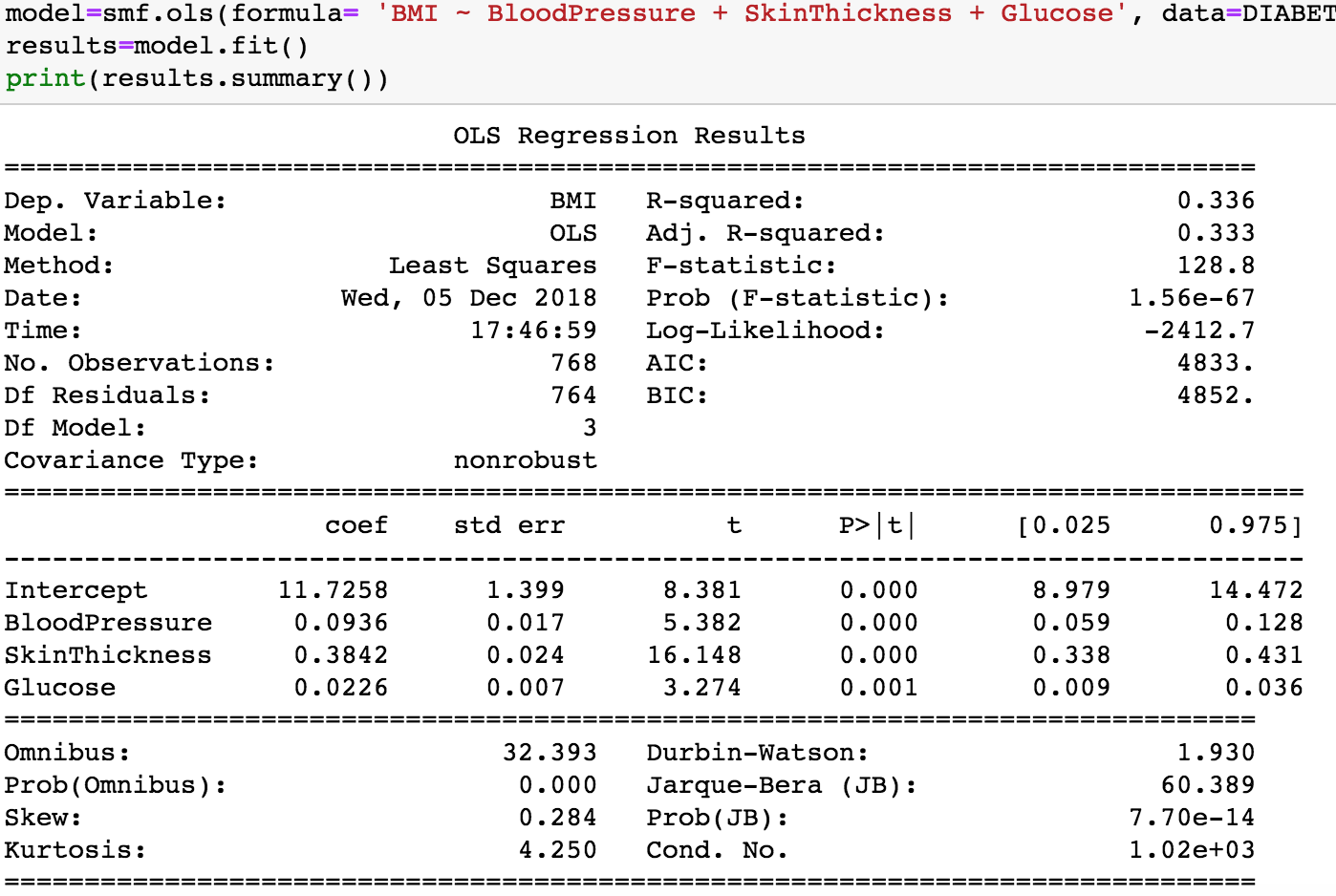








BMI Multiple Linear Regression:



**Classification**

To clarify the following data the 1 class is for people with diabetes and the 0 class is for people who do not have diabetes. The accuracy is for the diabetes class and when looking at the metrics like precision, F1 etc the first number x is for the diabetes class and the y is for the non diabetes class (x,y). The reason the data for our class 1 comes first for F1 score, precision etc. is because the confusion matrix was flipped and our diabetes class was being placed in class 0, but we flipped the matrix to put it in class 1 for visual purposes.

We performed various forms of classifications based on different features because we wanted to test whether a patient is diabetic or non-diabetic. We performed the classifications on the following:

1. Everything grouped together.
2. Highest difference in means grouped together (preg, glucose, Insulin, BMI, DPF, Age).
3. Lowest p-values grouped together (Glucose, BMI).
4. Highest R-Squared in multiple regression grouped together.

We began our classifications with Decision Trees, Naive Bayes, K Nearest Neighbor, and Linear SVM. After performing tests for all of the above four features, we came to the conclusion that Decision Trees was not the best form of classification to perform for our dataset since we did not have many rectilinear decision boundaries to split on. We then tried K-NN which was not the best method because it because it does not provide descriptive values. When we performed Naive Bayes and Linear SVM classification method we got better results in terms of accuracy, F1 Score, recall, precision. For Naive Bayes we preferred how it gave us a descriptive value for the class label and attributes, we also liked how it was applicable to categorical and numeric attributes, and how it is robust to noise. For linear SVM’s we liked how it was applicable to numeric and categorical attributes, and how it finds an optimal decision boundary.

After performing classifications for everything grouped together, our best form of classification was Linear SVM which gave us an accuracy of 0.78125, error of 0.21875, F1\_score of 0.8478 for class 1 and 0.6111 for class 0, and a precision of 0.801 for class 1 and 0.717 for class 0.

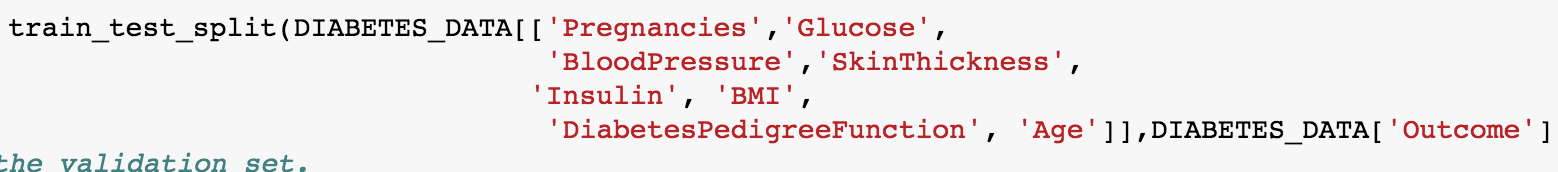
Our best classification for highest difference in means grouped together for ‘Pregnancies’, ‘Glucose’, ‘Insulin’, ’BMI’, ’DPF’, ’Age’ was Naive Bayes which gave us an accuracy of 0.7865, error of 0.2135, F1\_score of 0.8509 for class 1 and 0.6239 for class 0.

The best classification for lowest p-values grouped together ‘BMI’ and ‘Glucose’ was Linear SVM which gave us an accuracy of 0.78125, error of 0.21875, precision of 0.8014 for class 1 and 0.7174 for class 0 and an F1\_score of 0.8479 for class 1 and 0.6111 for class 0.

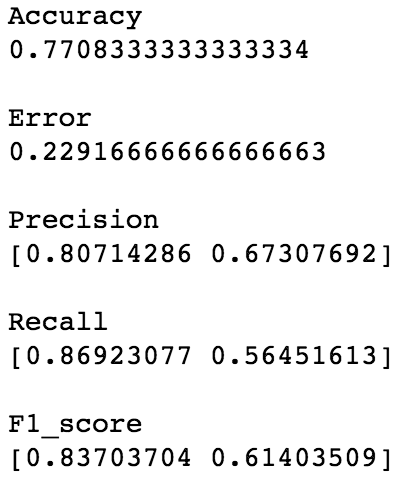
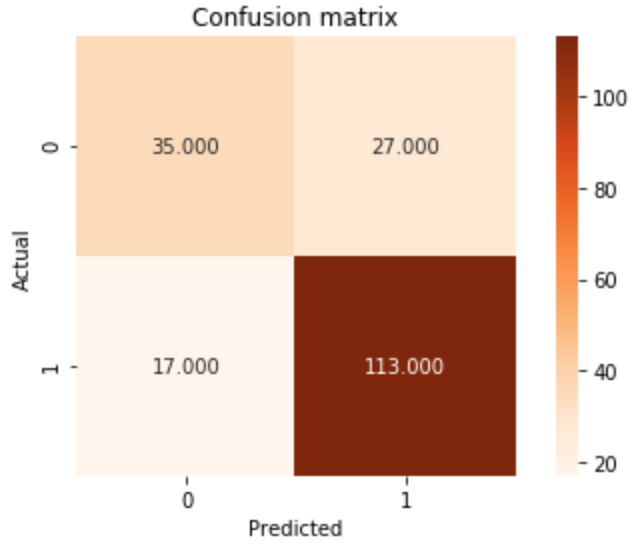
The highest R-squared in multiple regression grouped together for ‘BMI’, ‘Blood Pressure’, ‘Skin Thickness’, ‘Glucose’ was linear SVM which had an accuracy of 0.7865, error of 0.2135, precision of 0.8027 for class 1 and 0.7333 for class 0, and F1\_score of .8520 for class 1 and 0.6168 for class 0. For ‘BMI’, ‘Blood Pressure’, ‘Insulin’, ‘Glucose’, ‘Age’ our best form of classification was linear SVM which gave us an accuracy of 0.765625, error of 0.2344, precision of 0.79310 for class 1 and 0.68085 for class 0 and F1\_score of 0.8364 for class 1 and 0.5872 for class 0. For ‘BMI’, ‘Blood Pressure’, ‘Age’, ‘Glucose’ our best classification was linear SVM which gave an accuracy of 0.7760, error of 0.2240, F1\_score of 0.8436 for class 1 and 0.6055 for class 0 and precision of 0.8 for class 1 and 0.7021 for class 0.

The best model had a highest accuracy of 0.7865 and F1 score of 0.8520. The predictors used were from the multiple linear regression model: BMI, Blood Pressure, Skin Thickness, and Glucose. The relationship between BMI and the rest of these variables also lead to the highest R^2 value of 0.336, signifying a large correlation between BMI and the rest of the variables. After the Linear SVM classification we can tell that the correlation between BMI, Blood Pressure, Skin Thickness, and Glucose to predict the outcome of diabetes is both accurate, and has a high F1 score signifying a balanced relationship between precision and recall. We were able to increase the recall without decreasing precision. A lot of our models were good at predicting the data due to the intermediate steps we took in order to pick the best variables and best methods for classification.

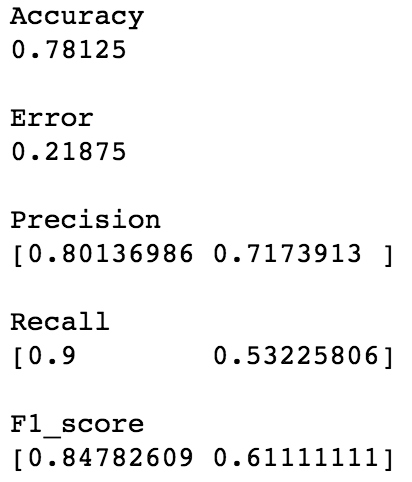
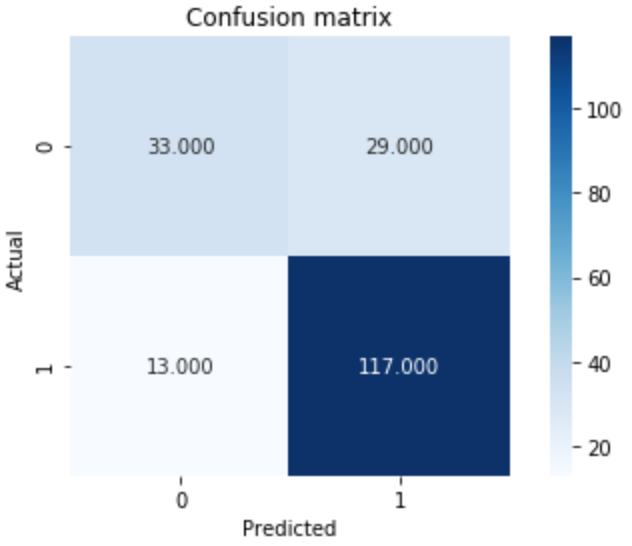
1. **Everything Grouped Together:**



Naive Bayes:

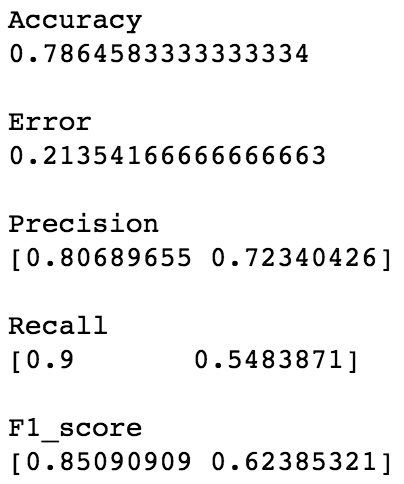
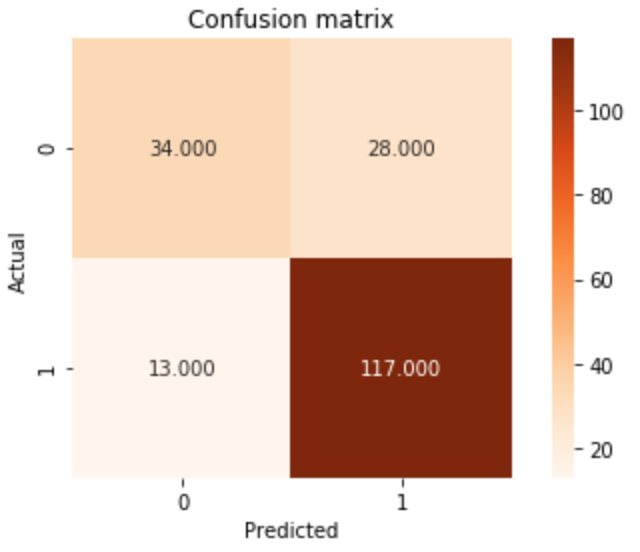


Linear SVM:

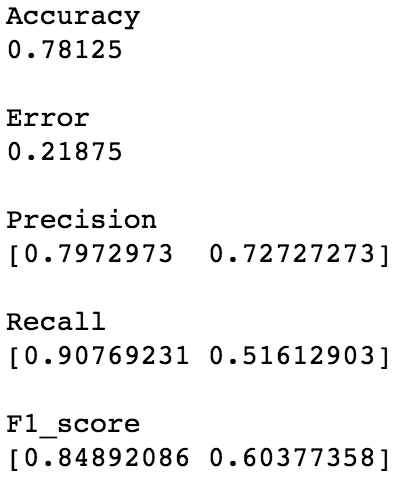
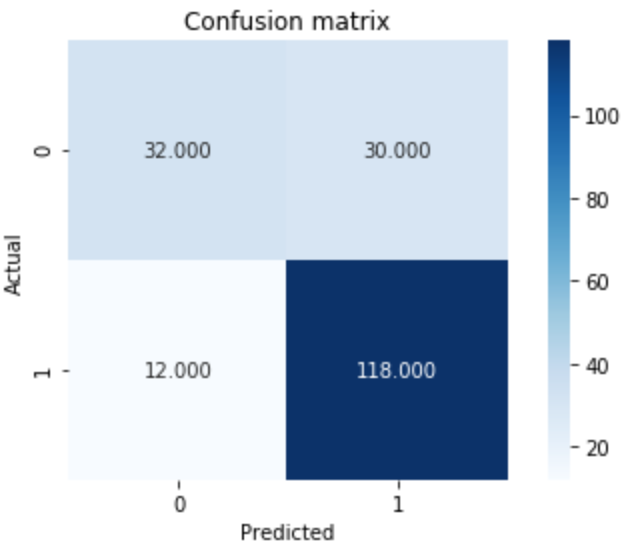


**2. Highest difference in means grouped together (preg, glucose, Insulin, BMI, DPF, Age)**

Naive Bayes:

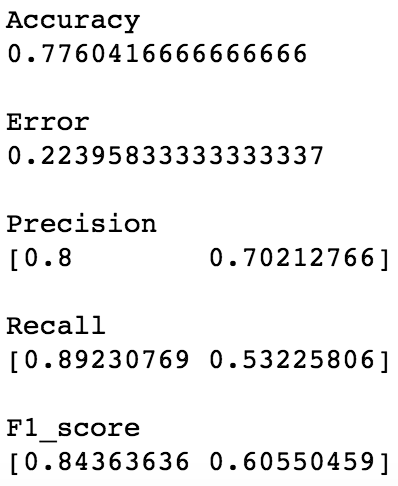
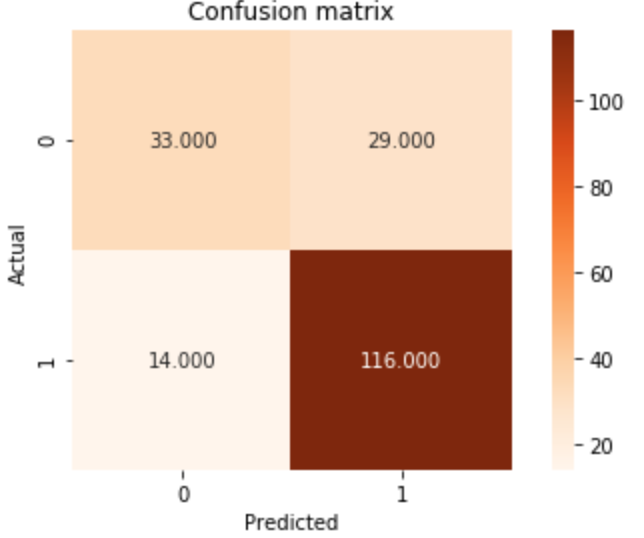
****

Linear SVM:

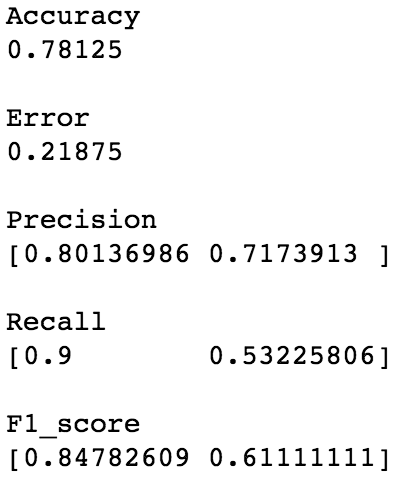
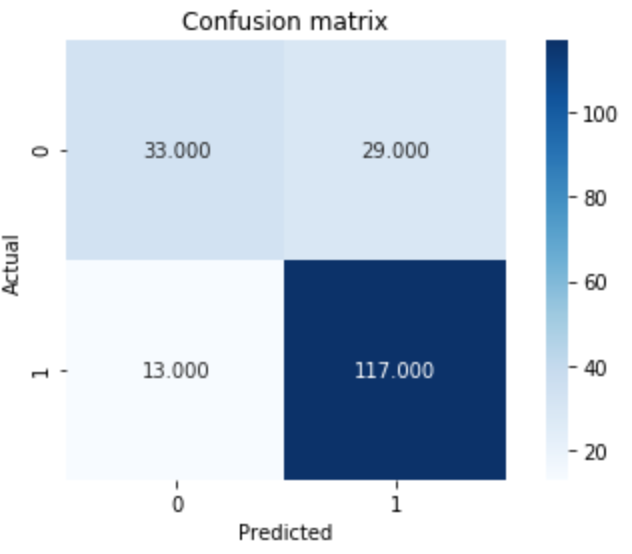


**3. Lowest p-values grouped together (Glucose, BMI)**

Naive Bayes



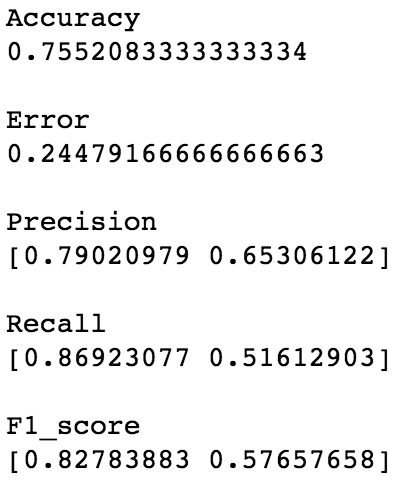
Linear SVM



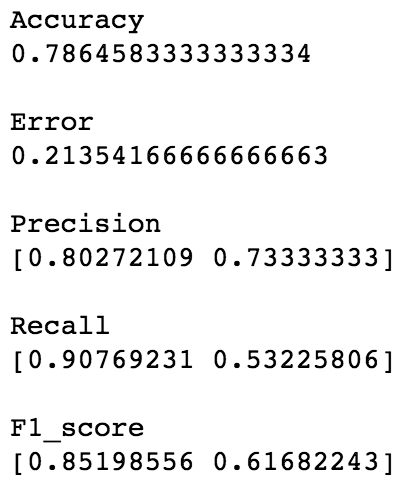
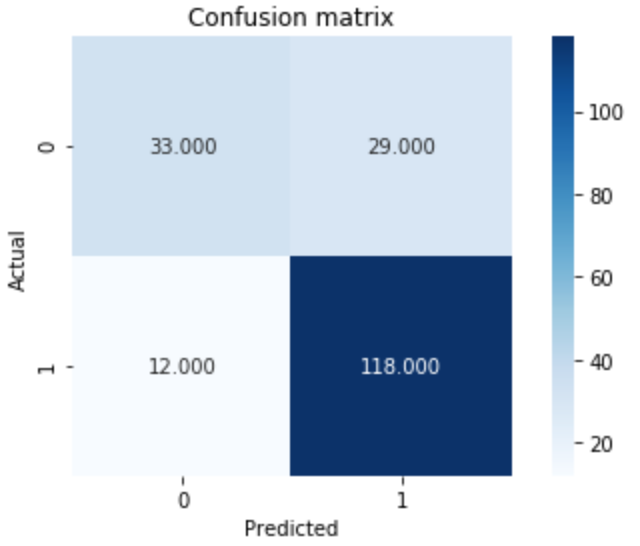
**4. Highest R-Squared in multiple regression grouped together**

****

Naive Bayes:

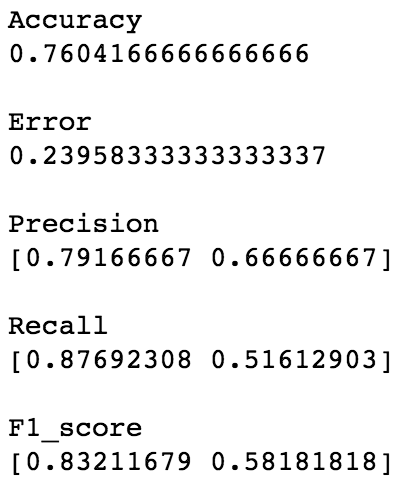
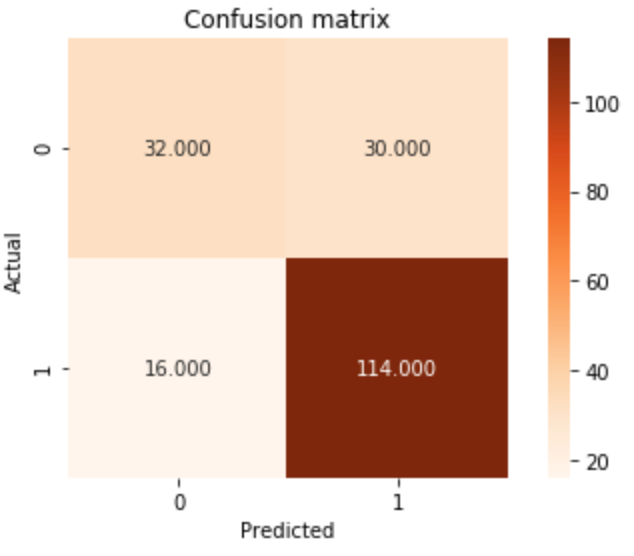


Linear SVM:

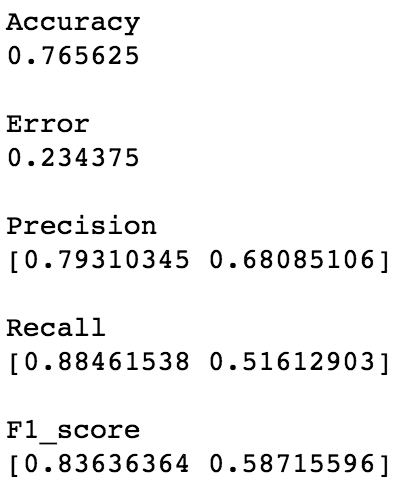




Naive Bayes:

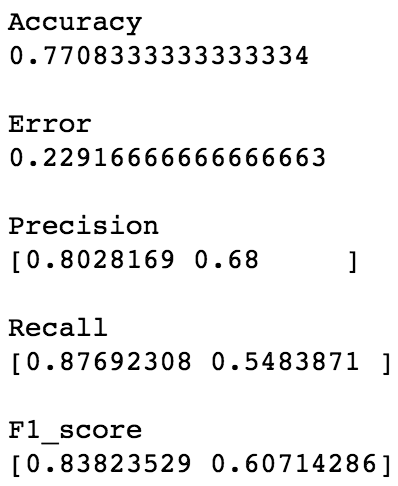
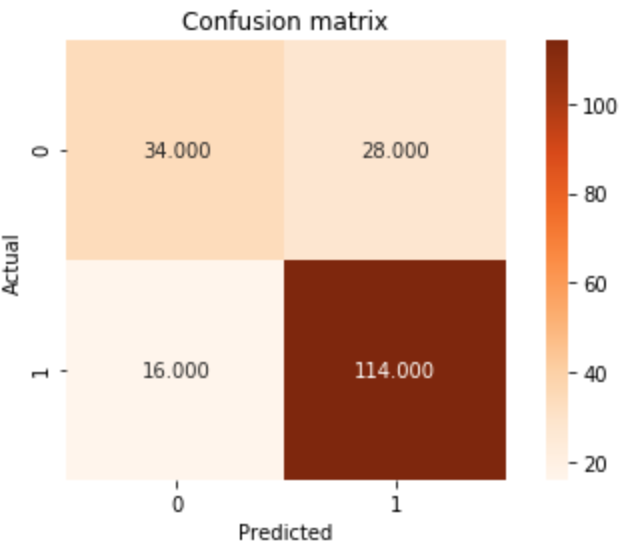


Linear SVM:

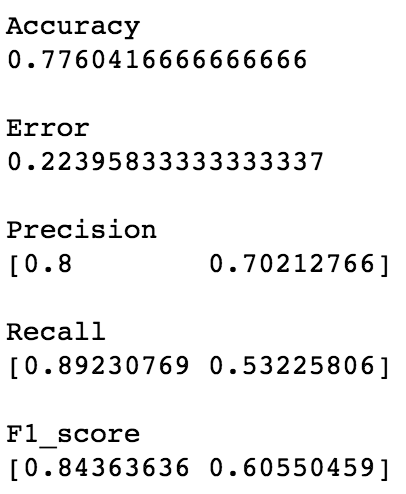
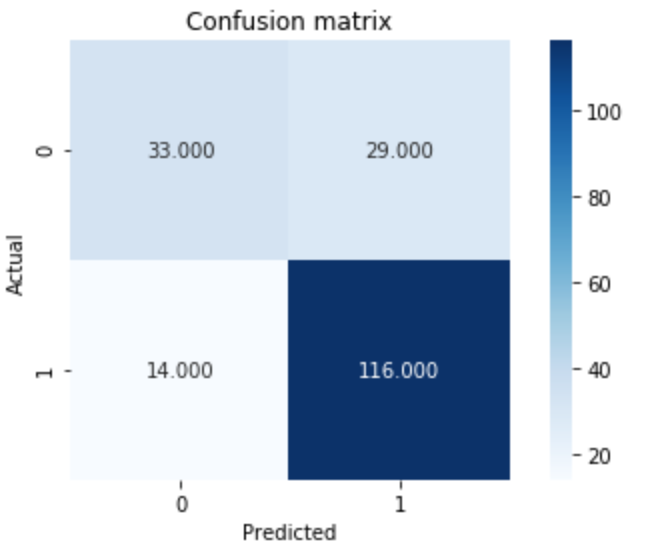




Naive Bayes:



Linear SVM:



**Resources**

1. <https://www.cdc.gov/media/releases/2017/p0718-diabetes-report.html>
2. https://www.washingtonpost.com/archive/lifestyle/wellness/1993/03/30/why-are-the-pima-indians-sick-studies-on-arizona-tribe-show-excessive-rates-of-diabetes-obesity-and-kidney-disease/1f978958-e73b-483a-9af9-47d9efdad534/?utm\_term=.f65b50af2565