Pima Indian Diabetes Dataset Analysis

Josh Cubero

Colleague ID: 0602801

MSBA 320

Golden Gate University

Abstract

The US House of Representatives recently passed a bill that would aid in capping the cost of insulin to $35 per month, which at present costs an estimated $1200 per month. Insulin is a medication that people suffering from diabetes must have to remain alive. According to the International Diabetes Federation, it’s estimated that 537M adults have diabetes. Additionally, IDF estimates that 240M of those with diabetes are undiagnosed. Lastly, the IDF states that diabetes caused 6.7M deaths and costs $966B in treatment costs.

Diabetes is a group of diseases that affect the body’s ability to either produce or use insulin (DiabetesResearch.org, 2022). Insulin is a hormone produced in the body’s pancreas and assists the body with processing glucose from blood and creating energy. If this glucose transformation process is disrupted, the body doesn’t process out the glucose, which in turn causes the blood sugar to increase. Individuals whose blood sugar remains elevated for an extended period are at risk for serious medical problems, perhaps fatal.

There are two types of diabetes, Type 1, and Type 2. Type 1 diabetes is an autoimmune condition where the person’s immune system attacks cells in the pancreas where insulin is produced. Once a significant number of pancreatic cells are destroyed, the body is no longer able to produce insulin to regulate glucose levels. According to diabetesresearcher.org, researchers are not certain what causes Type 1 diabetes, as the onset of type 1 can occur suddenly and without warning. Type 2 diabetes is a condition in which the pancreas continues producing insulin, however, the body rejects it. With time, as the body rejects insulin, the insulin gradually becomes ineffective, and glucose accumulates in the bloodstream.

Doctors believe there are risk factors that contribute to the onset of type 2 diabetes. According to heart.org, there are two types of risk factors for type 2 diabetes: Non-modifiable

and modifiable. The non-modifiable risk factors are those the person can't change such as family medical history, race or ethnicity, age, and gestational diabetes which develops during pregnancy. Diabetes type 2 modifiable risk factors are those that the person can change by modifying their lifestyle. The first is the person's bodyweight, as suffering from obesity or being overweight increases the risk of developing type 2 diabetes. Next, there's physical activity levels, which aid in reducing the body's resistance to insulin (Heart.org, 2022). In addition, the previous two risk factors, individuals with high blood pressure, high cholesterol, smoking, poor dietary habits, excessive alcohol use, poor stress management, and inefficient sleep could be at risk for type 2 diabetes.

Fortunately, researchers and doctors believe type 2 diabetes is preventable. According to the Mayo Clinic, people can reduce the risk of developing type 2 diabetes by losing weight, being physically active, having a balanced diet rich in plant-based foods, consuming healthy fats, and avoiding fad or crash diets. For years doctors have been able to run tests, observe the results, and based on experience, mathematics, then state whether the person is at risk for type 2 diabetes.

In today's world of technology and machines, is there a way to perform medical tests, then input the results to a machine that could predict the onset of type 2 diabetes? That is the question this study will be founded on. The purpose of this study is to perform exploratory data analysis on the Pima Indian Type 2 Diabetes dataset, perform ANOVA testing, then create predictive model based on logistic regression and supervised machine learning. The results of the machine learning model are not intended to replace the advice of a medical professional but are simply a test for proof of concept.

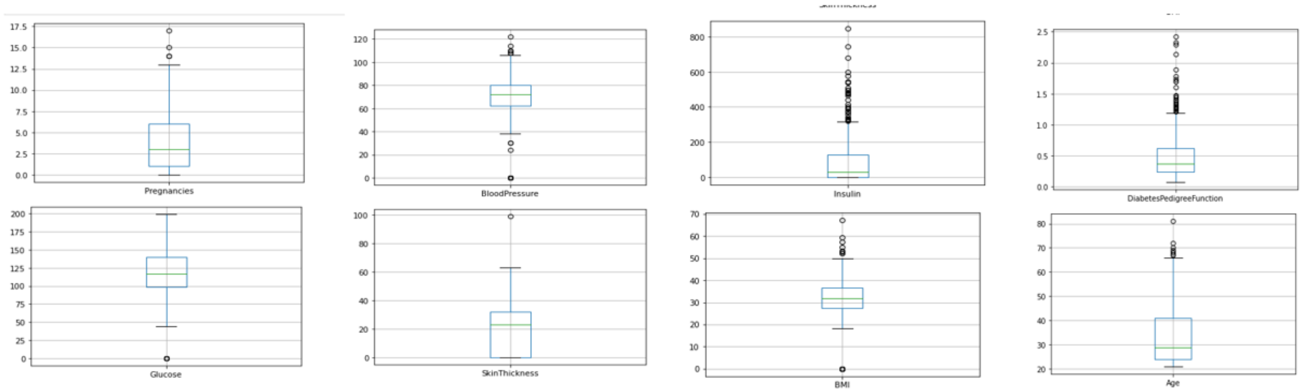
Gathering data

The data for this study was originally created by the National Institute of Diabetes Digestive and Kidney Diseases. The database was donated by Vincent Sigillito of the Research Center at Johns Hopkins University in May 1990. This dataset was used by a team of researchers in 1988 to conduct a diagnostic analysis. The team was able to build a predictive model using the split model method. They created a train and test dataset and built an algorithm that was 76% accurate (Kaggle, 2022). In this study, we shall endeavor to meet or exceed the 1988 study's probability of accuracy.

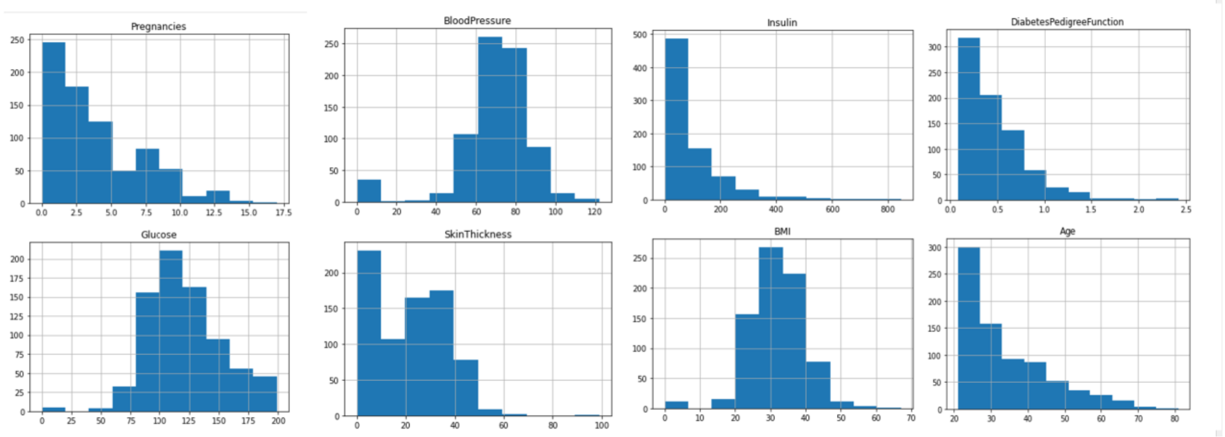
The data was collected and condensed prior to its posting on Kaggle.com. The data posted was filtered to include only female patients at least 21 years of age and of Pima Indian descent. Variables in the dataset are Number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index, Diabetes Pedigree Function, Age, and Outcome, which is the response variable. All explanatory variables are of numeric type, while the response variable is binary.

Descriptive Statistic

The basis for the descriptive statistics portion of the study will be explained by evaluating the boxplots for each of the explanatory variables. A look at each boxplot and we can see that all the explanatory variables have outliers. Two of the explanatory variables, Insulin and Diabetes Pedigree Function, have a significant number of outliers. Another item of note is the difference in scale between the explanatory variables, with insulin's max at 800 and diabetes pedigree function's max at just 2.5.

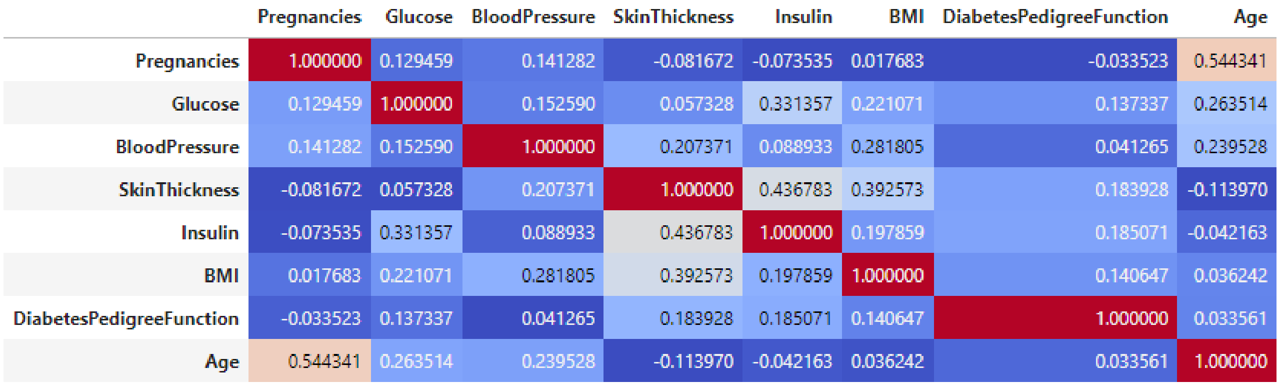


Next, each explanatory variable is plotted on a histogram to observe their distributions and the effects outliers on variables with a high count of outliers. If one observes the Insulin, Diabetes Pedigree Function, and Age, one will see that these three variables are heavily skewed to the right. Additionally, the explanatory variables do not appear to be normally distributed. This will likely affect the results of an ANOVA test since ANOVA assumes the data is normally distributed and has homogeneous variance (Penn State, 2022).



Correlation between the variables

The next item to discuss is the correlation between the independent variables. Observing correlation between independent variables is important because of multicollinearity. Multicollinearity exists when independent variables have at least a moderate correlation, which increases covariance estimates and causes the variance to become inflated (Penn State, 2022).

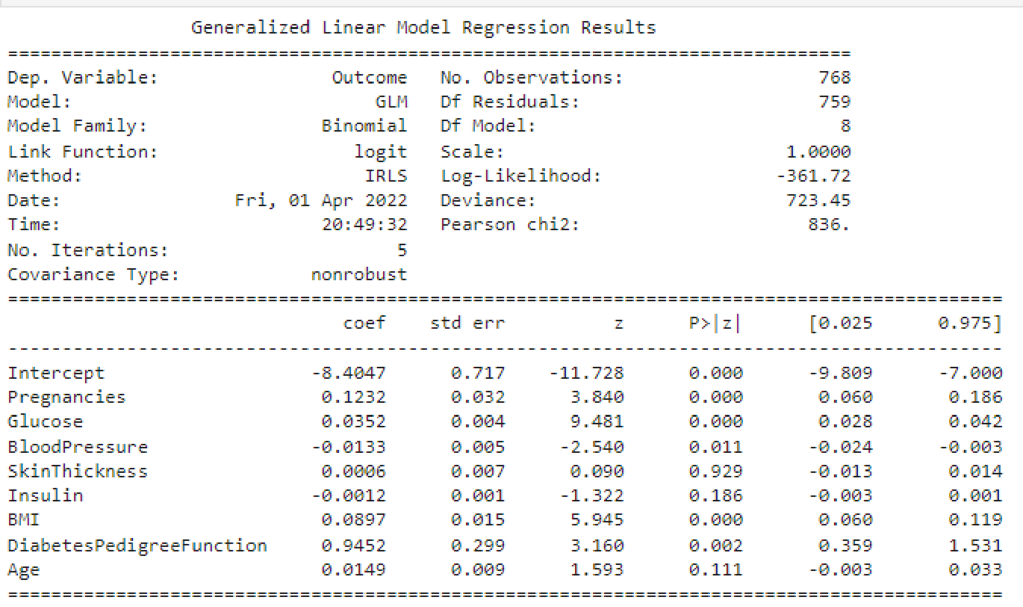


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000

The figure above depicts the heatmap of independent variables created to search for the presence of collinearity. The greatest collinearity in the dataset exists between age and pregnancies. This existence appears logical as the number of pregnancies could be more likely to increase as the person grew in age. Based on the correlation matrix, multicollinearity shouldn't be a factor in this study.

Logistic Regression

Now that descriptive statistics and correlation have been analyzed, the study will evaluate logistic regression. As observed in the visual analysis of the independent variables, the dataset is not normally distributed. Additionally, the response variable is binary, which means that performing a linear regression would not likely yield the correct results. This study used the generalized linear model, with binomial selected as family and logit selected as the link function. The following are the results of the logistic regression.

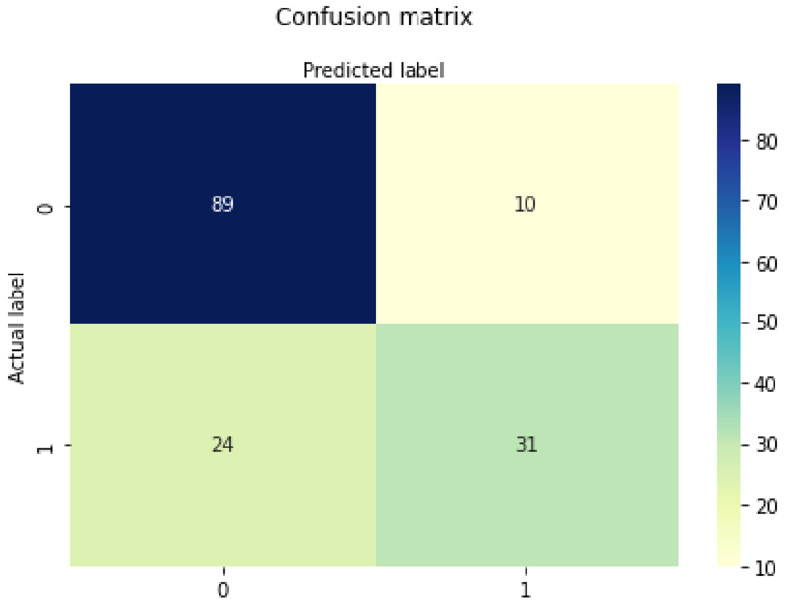


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  768
Model:                            GLM   Df Residuals:                      759
Model Family:                Binomial   Df Model:                            8
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -361.72
Date:                Fri, 01 Apr 2022   Deviance:                       723.45
Time:                        20:49:32   Pearson chi2:                     836.
No. Iterations:                     5
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   -8.4047      0.717    -11.728      0.000      -9.809      -7.000
Pregnancies                  0.1232      0.032      3.840      0.000       0.060       0.186
Glucose                      0.0352      0.004      9.481      0.000       0.028       0.042
BloodPressure               -0.0133      0.005     -2.540      0.011      -0.024      -0.003
SkinThickness                0.0006      0.007      0.090      0.929      -0.013       0.014
Insulin                     -0.0012      0.001     -1.322      0.186      -0.003       0.001
BMI                          0.0897      0.015      5.945      0.000       0.060       0.119
DiabetesPedigreeFunction     0.9452      0.299      3.160      0.002       0.359       1.531
Age                          0.0149      0.009      1.593      0.111      -0.003       0.033
============================================================================================
```

The results of the linear regression indicate that there are three independent variables that are not significant: Skin Thickness, Insulin, and Age. Additionally, the model's deviance is 723.45, which is an indicator that the model is not a perfect fit. In a logistic regression, deviance is like sum of squares, and is an indicator of the model's goodness of fit. Deviance in logistic regression can range from 0 to infinity, so a deviance of 723.45 doesn't necessarily mean the model will perform poorly.

Predictive Modeling

The final analysis conducted in this study hopes to be the most impactful by building a predictive model to predict the outcome based on input variables. The first step was to build training and testing x-y variables. Once these variables were instantiated, a logistic regression variable was created and used to fit the X and Y training data. The results of this model are depicted in the confusion matrix below.



The confusion matrix is divided into four quadrants that depict the outcomes of the predictive model. The four quadrants in this study can be explained as follows:

- True Positive: A true positive result is one where the actual result was positive, and the predicted result was positive (Vitalflux, 2021). In this model there were 31 instances of true positive outcomes.
- False Positive: This result occurs when the model predicts a negative result, however the actual result was positive (Vitalflux, 2021). In this model there were 24 false positives.
- True Negative: The true negative result occurs when the model predicts a negative result, and the outcome is negative as well (Vitalflux, 2021). In the case of this model, there were 89 true negative outcomes.
- False Negative: A false negative occurs when the prediction is positive, but the actual is negative (Vitalflux, 2021). In this model there were 10 false negative outcomes.

The final performance evaluation for the predictive model is to review Accuracy, Precision, and Recall. The accuracy score is the ratio of true positives plus true negatives and the sum of all four possible outcomes. The accuracy metric essentially indicates the likelihood of the model predicting a correct outcome given the total number of predications made. Next, precision score is the ratio of true positives to the sum of true and false positive predictions made. Finally, recall is the ratio of true positives to all positive outcomes, whether true or false positive (Vitalflux, 2021). The accuracy, precision, and recall for this model is as follows: Accuracy: 0.78, Precision: 0.76, Recall: 0.56.

Results

The results of this study indicate that the predictive model has 78% accuracy rate. However, the recall metric is lower than the model's overall accuracy, which indicates the model is not efficient at accurately predicting positive outcomes. This study proves that medical data can be used to create a predictive model using logistic regression. However, any predictive model intended for medical use should explore and compare using other machine learning techniques such as K Nearest Neighbor or Random Forest.

Conclusion

In conclusion, this paper has validated that it is plausible to take results from medical samples such as those in this dataset to predict whether a person could develop diabetes. The reader is encouraged to trust their doctors and medical professionals when warned of the risks of developing type 2 diabetes, as the logic behind the warnings has been proven in this study.

References

Types of Diabetes | Diabetes Research. (n.d.). Www.diabetesresearch.org. https://www.diabetesresearch.org/types-of-diabetes

American Heart Association. (2015). *Understand Your Risk for Diabetes*. Www.heart.org. https://www.heart.org/en/health-topics/diabetes/understand-your-risk-for-diabetes

Mayo Clinic. (2019). *Diabetes prevention: 5 tips for taking control*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/in-depth/diabetes-prevention/art-20047639

Mehmet Akturk. (2020). *Diabetes Dataset*. Kaggle.com. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

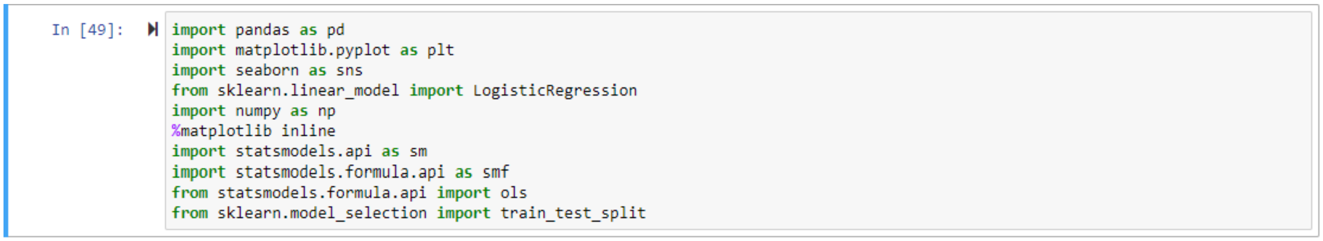
Facts & figures. (n.d.). Www.idf.org. Retrieved April 2, 2022, from https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html#:~:text=Diabetes%20facts%20%26%20figures%201%20Approximately%20537%20million

Accuracy, Precision, Recall & F1-Score - Python Examples. (2021, October 1). Data Analytics. https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

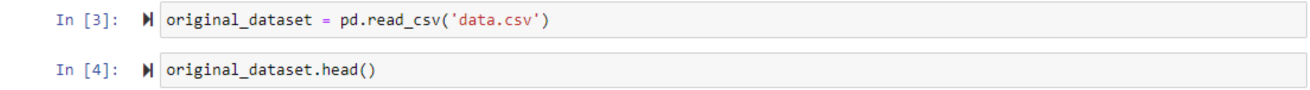
10.2.1 - ANOVA Assumptions | STAT 500. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat500/lesson/10/10.2/10.2.1

10.4 - Multicollinearity | STAT 462. (n.d.). Online.stat.psu.edu. https://online.stat.psu.edu/stat462/node/177/

Coding

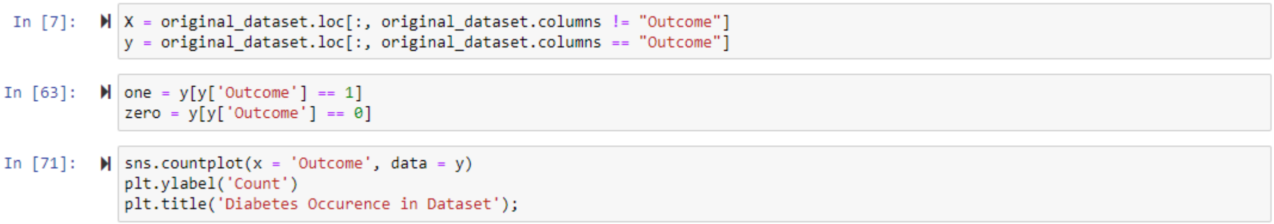


In [49]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
import numpy as np
%matplotlib inline
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols
from sklearn.model_selection import train_test_split
```



In [3]:
```
original_dataset = pd.read_csv('data.csv')
```

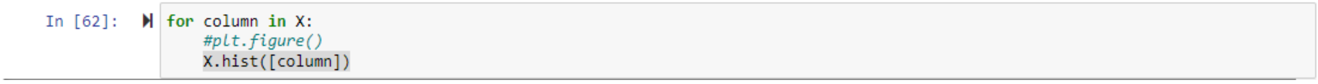
In [4]:
```
original_dataset.head()
```



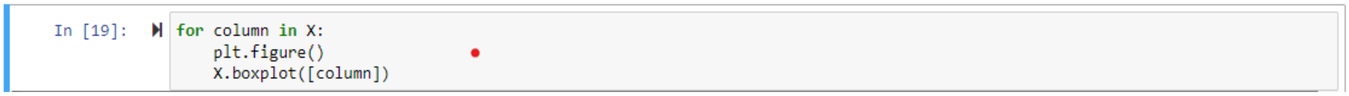
In [7]:
```
X = original_dataset.loc[:, original_dataset.columns != "Outcome"]
y = original_dataset.loc[:, original_dataset.columns == "Outcome"]
```

In [63]:
```
one = y[y['Outcome'] == 1]
zero = y[y['Outcome'] == 0]
```

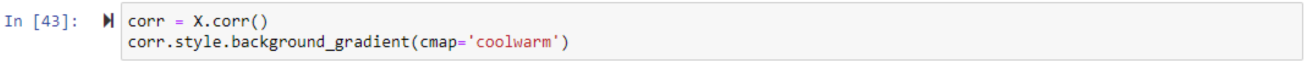
In [71]:
```
sns.countplot(x = 'Outcome', data = y)
plt.ylabel('Count')
plt.title('Diabetes Occurence in Dataset');
```



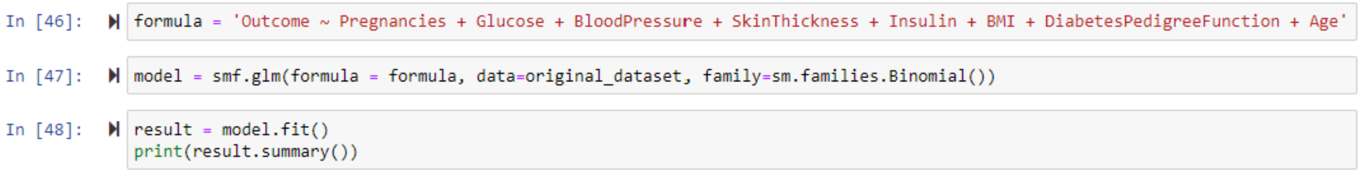
In [62]:
```
for column in X:
    #plt.figure()
    X.hist([column])
```



In [19]:
```
for column in X:
    plt.figure()
    X.boxplot([column])
```



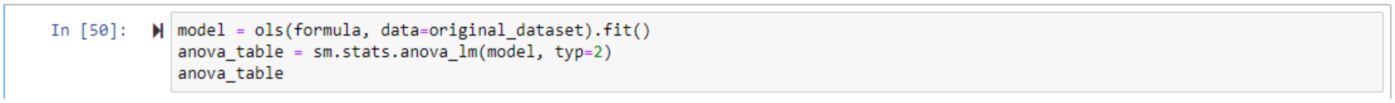
In [43]:
```
corr = X.corr()
corr.style.background_gradient(cmap='coolwarm')
```



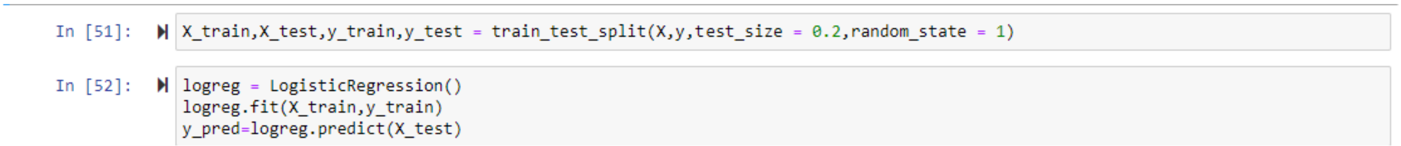
In [46]:
```
formula = 'Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFunction + Age'
```

In [47]:
```
model = smf.glm(formula = formula, data=original_dataset, family=sm.families.Binomial())
```

In [48]:
```
result = model.fit()
print(result.summary())
```



In [50]:
```
model = ols(formula, data=original_dataset).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```



In [51]:
```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.2,random_state = 1)
```

In [52]:
```
logreg = LogisticRegression()
logreg.fit(X_train,y_train)
y_pred=logreg.predict(X_test)
```

In [54]:

```
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

Out[54]:

```
array([[89, 10],
       [24, 31]], dtype=int64)
```

In [55]:

```
class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

0 1

In [56]:

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```