Predicting Diabetes using Logistic Regression

C4 Team Members:

Harshil Surendra Chauhan

Neha Singh

Rachel Sherry Sunder Raj

Ramya Datla

1. Abstract:

By Ramya Datla

The data has been collected from "National Institute of Diabetes and Digestive and Kidney Diseases" as part of the Pima Indians Diabetes Database. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. The number of patients in the database n=768 each with 9 attribute variables. Out of the nine conditional attributes, six are due to physical examination rest of the attributes are chemical examination. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of machine learning technique. This project aims to predict diabetes via supervised machine learning methods Logistic regression. This project also aims to propose an effective technique for earlier detection of the diabetes disease. The datasets include data from 768 women with several medical predictor variables and one target variable. The classification goal is to predict whether the patients in the dataset have diabetes or not.

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. This is used for predicting the categorical dependent variable, using a given set of independent variables. It predicts the output of a categorical variable, which is discrete in nature. Moreover, it is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes) or 0 (no). This type of analysis can help you predict the likelihood of an event happening or a choice being made. When using machine learning algorithms, we should always split our data into a training set and test set. In our case, we will also separate out some data for manual cross checking. As the final step before using machine learning, we will normalize our inputs. We are presenting a proposal about predicting diabetes using Regression Model. This model approach for clinical analysis that is a probabilistic estimation which helps in understanding the relationship between the dependent variable and one or more independent variables.

2. Introduction:

By Neha Singh

Here we are presenting a proposal about predicting diabetes using regression. It is used in this model approach for clinical analysis that is a probabilistic estimation which helps in understanding the relationship between the dependent variable and one or more independent variables. Diabetes, being one of the most common diseases around the world when detected early may prevent the progression of the disease and avoid other complications. Here, we design a prediction model that predicts whether a patient has diabetes, based on certain diagnostic measurements included in the dataset, and explore various techniques to boost the performance and accuracy. The classification goal is to predict whether the patients in the dataset have diabetes or not. To run the analysis, we import pandas to peruse our information from a CSV document and control it for additional utilization. Further, utilizing NumPy to change over out information into an organization appropriate to take care of our order model.

2.1 Dataset & Model Approach

The dataset found 700+ which is large enough for us to do the logistic regression as it performs well when the data is discrete and has good accuracy for many simple data sets. Logistic regression is more straightforward to apply, analyze, and train.

Table 1: Sample Dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30
3	78	50	32	88	31	0.248	26
10	115	0	0	0	35.3	0.134	29
2	197	70	45	543	30.5	0.158	53
8	125	96	0	0	0	0.232	54
4	110	92	0	0	37.6	0.191	30
10	168	74	0	0	38	0.537	34
10	139	80	0	0	27.1	1.441	57

2.2 Data Cleaning

The process of detecting and correcting the corrupt or incorrect records from a document set, table, or database is known as data cleansing, and it entails identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. And then the data were loaded to slit it and check the accuracy of it, then train and at last tested it.

2.3 Measurements for Model Comparison

Mean Squared Error (MSE): It is calculated as MSE = 1 $n\sum$ (yi − $\hat{y}$) 2 n i=1 , if n is the number of observations, yi the real value of the target variable and $\hat{y}i$ the predicted value from the model (Tiryaki, 2014). As this metric squares the difference between prediction and real value, it penalizes errors of higher absolute difference compared to MAE. The lower this value is, the better.

Bayesian information criterion (BIC) : It follows the same logic as AIC, only that the penalty term is weighted by $log\ n$ instead of 2, with n being the number of observations. Again, the lower BIC is, the better the model. BIC tends to choose fewer variables than AIC (Kuha, 2004). To this analysis, the RMSE is chosen. RMSE is a very popular measurement as it retains the units of the target variables and therefore, interpretation of results is more intuitive. Looking at the MSE would yield in the same ranking of models, as $RMSE$ = √MSE.

2.4 Model selection

Ordinary Least Squares (OLS): In OLS, the objective is to minimize the residual sum of squares $SSres$ = ∑ (yi − $\hat{y}$) 2 n i=1 . In other words, the estimator $\boldsymbol{\beta}$ = $arg\ min\ \beta$ ‖$\boldsymbol{y}$ − $\boldsymbol{X\beta}$‖ is chosen so that the Euclidean distance between prediction and actual values are minimized. In this formulation, $\boldsymbol{\beta}$ is a vector of dimensionality equal to the number of independent variables plus 1, as it includes one estimator for each variable and one for the intercept β 0. $\boldsymbol{X}$ denotes the set of datapoints in the model. The final regression line then is given by $\hat{\boldsymbol{y}}$ = β 0 + β 1$\boldsymbol{x}$1 + β 2$\boldsymbol{x}$2 + ⋯ + β $d\boldsymbol{x}d$, where the bold face y and x are vector representations of the predicted values and observations for variable 1 to d, respectively, meaning the dependent variable is in a linear relationship to the predictor variables (Pohlman, Leitner, 2003). As for other assumptions of OLS

regressions, the error terms are independent from each other and follow a normal distribution with mean 0 and fixed standard deviation σ.

Lasso: Here, instead of the sum over βj 2, the penalty term is identified as $\lambda \sum |\beta j| \; p \; j$=1. This results in some of the coefficients to shrink to exactly zero, instead of just getting close to it (Ogutu et al., 2012). Therefore, given a large enough λ, Lasso performs a variable selection process. Again, cross validation is used to determine the best value for λ. The assumptions of lasso regression follow those of OLS regression models, though the assumption regarding the distribution of the error term is weaker (i.e., lasso can handle different distributions as well).

Principal Component Analysis (PCA): In PCA, the goal is to project the dataset of dimensions $n \times p$ onto a smaller dimension, say $n \times l$, where n is the number of observations and p the number of independent variables in the dataset, with $1 \leq l \leq p$. In each iteration, OLS is used to create a linear model with the newly built set of components. Finally, the model with the lowest RMSE is chosen for model comparison. Hence, each of the datasets can have an individual set of principal components. PCA helps to model datasets that suffer from high collinearity. As mentioned before, the underlying dataset contains multiple variables that share high correlations. The results will show that this causes a high collinearity coefficient. Therefore, this method poses as a promising approach.

3. Results:

By Rachel Sherry

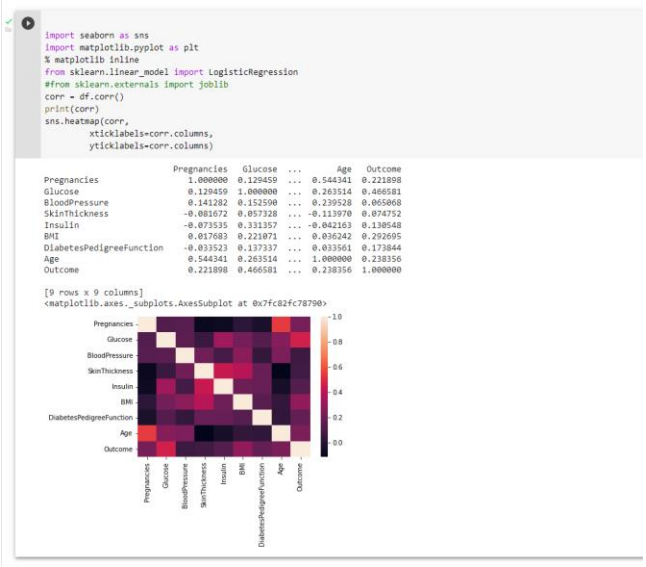
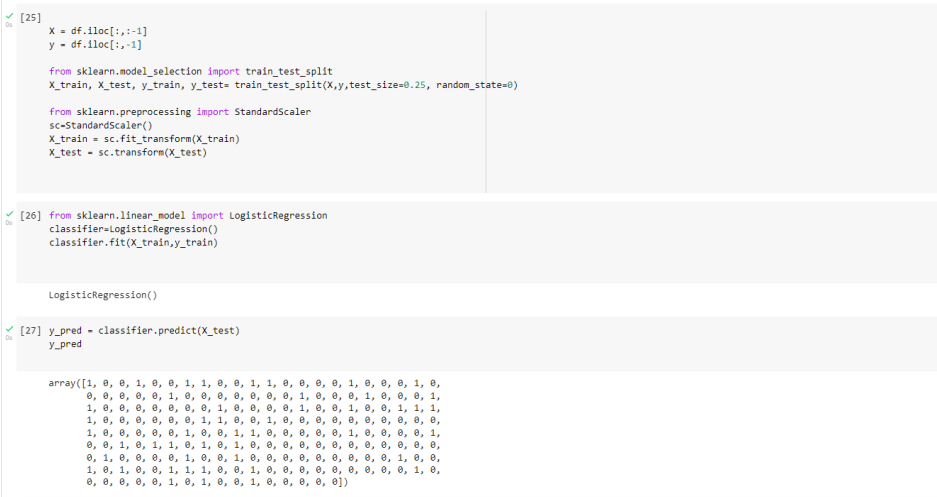


Figure 3.1: Heatmap of data

To begin exploring our data we first get insights about it by obtaining the correlations of every pair of features with respect to the outcome variable. We visualise this with a heatmap that is shown in Figure 3.1, brighter colours indicate more correlation. We see that the glucose levels, age, BMI, and number of pregnancies all have significant correlation with the outcome variable.

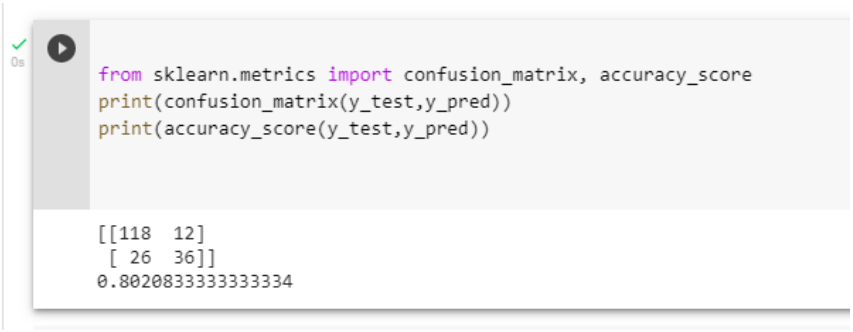
3.1 Training and Testing data



```
[25]
X = df.iloc[:,:-1]
y = df.iloc[:,-1]

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test= train_test_split(X,y,test_size=0.25, random_state=0)

from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

[26] from sklearn.linear_model import LogisticRegression
classifier=LogisticRegression()
classifier.fit(X_train,y_train)

LogisticRegression()

[27] y_pred = classifier.predict(X_test)
y_pred

array([1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
       0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0])
```

Figure 3.2: Training and Testing data

In our project we have divided our data into two sets to train and test them. We now train our classification model by using logistic regression, for this we use *sklearn*. First, we create an instance and then use the fit function to train the model. The data is split into training and testing sets, the data is split in the ratio of 70:30, i.e., 70% for training and 30% for testing. A Logistic Regression algorithm is used to make the predictions and check for the accuracy.



```
from sklearn.metrics import confusion_matrix, accuracy_score
print(confusion_matrix(y_test,y_pred))
print(accuracy_score(y_test,y_pred))

[[118  12]
 [ 26  36]]
0.8020833333333334
```

Figure 3.3: Accuracy achieved and Confusion Matrix

The execution time is also calculated. For predictions, there are four important terms i.e., True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP and TN represent

the cases when the actual outcome and the result are the same, whereas FP and FN are the cases when the opposite results are obtained. A classification report is generated which includes Precision, Recall, F1 score, and Support. The Precision metric shows what percent of predictions are correct. The training and testing are again carried out using these new features, and the accuracy and execution time are noted. Addition of new features can observe an increase in the accuracy and a major change in the runtime of the program, which decreased to a great extent. This highlights the importance of feature selection in improving model performance.

3.3 OLS Regression

```
[31] #OLS Regression Model

     from sklearn import datasets
     import statsmodels.api as sm
     import statsmodels.formula.api as smf
     import pandas as pd
     import numpy as np
     from itertools import cycle
     from tqdm.auto import tqdm

     model = sm.OLS(y,X).fit()
     pred = model.predict(X)
     pred
     model.summary()
```

OLS Regression Results

Dep. Variable:	Outcome	**R-squared (uncentered):**	0.487
Model:	OLS	**Adj. R-squared (uncentered):**	0.481
Method:	Least Squares	**F-statistic:**	90.10
Date:	Mon, 13 Dec 2021	**Prob (F-statistic):**	9.04e-105
Time:	18:18:03	**Log-Likelihood:**	-429.34
No. Observations:	768	**AIC:**	874.7
Df Residuals:	760	**BIC:**	911.8
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Pregnancies	0.0247	0.005	4.536	0.000	0.014	0.035
Glucose	0.0039	0.001	7.701	0.000	0.003	0.005
BloodPressure	-0.0049	0.001	-6.005	0.000	-0.007	-0.003
SkinThickness	5.752e-05	0.001	0.049	0.961	-0.002	0.002
Insulin	4.783e-05	0.000	0.304	0.761	-0.000	0.000
BMI	0.0042	0.002	2.094	0.037	0.000	0.008
DiabetesPedigreeFunction	0.0980	0.048	2.059	0.040	0.005	0.191
Age	-0.0010	0.002	-0.652	0.515	-0.004	0.002

Omnibus:	294.015	**Durbin-Watson:**	2.034
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	80.012
Skew:	0.574	**Prob(JB):**	4.22e-18
Kurtosis:	1.913	**Cond. No.**	575.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 3.4: OLS Regression Results

The method of Ordinary Least Squares(OLS) is most widely used model due to its efficiency. This model gives best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑ei2). Dependent variable is one that is going to depend on other variables. In this regression analysis Y is our dependent variable which predicts the diabetic

condition of an input because we want to analyse the effect of X on Y. The number of observations is the size of our sample, i.e., N = 768. In regression we omit some independent variables that do not have much impact on the dependent variable, the intercept tells the average value of these omitted variables and noise present in model. The coefficient term tells the change in Y for a unit change in X. There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. While calculating p values we rejected the null hypothesis we can see same in C.I as well. R2 is the coefficient of determination that tells us that how much percentage variation independent variable can be explained by independent variable. Here, 48.7 % variation in Y can be explained by X. The maximum possible value of R2 can be 1, means the larger the R2 value better the regression. F – statistic tells the goodness of fit of a regression; it is noted to be F>90.10 for 1 and 8 data frame model.

3.4 Analysis

```
Model with rank: 1
Mean validation score: 0.827 (std: 0.040)
Parameters: {'lg__C': 0.1, 'pca__n_components': 8}

Model with rank: 2
Mean validation score: 0.826 (std: 0.040)
Parameters: {'lg__C': 0.1, 'pca__n_components': 7}

Model with rank: 3
Mean validation score: 0.826 (std: 0.040)
Parameters: {'lg__C': 0.1, 'pca__n_components': 6}
```

Figure 3.5: Principal Component Analysis

In principal component analysis, this relationship is quantified by finding a list of the principal axes in the data and using those axes to describe the dataset. The model with rank1 has a value of *0.827* that is noted to be the highest. To fit a lasso model, we use the Lasso() function. We have included the constraints to max_iter = 1000 to fit a ridge model. This is substantially lower than the test set MSE of the null model and of least squares, and only a little worse than the test MSE of ridge regression with alpha chosen by cross-validation. However, the lasso has a substantial advantage over ridge regression in that the resulting coefficient estimates are sparse. The best model is selected by cross-validation.

```
[ 0.02059187  0.00592027 -0.00233188  0.00015452 -0.00018053  0.01324403
  0.14723744  0.00262139]
['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']


Selected Variables MSE: 0.15982225092680033
All Variables MSE: 0.15982225092680033
LassoCV MSE: 0.15996699044658474
```

Figure 3.6: Comparing their test-set performance using MSE

3.6 AUC model results

ROC curves in logistic regression are used for determining the best cut-off value for predicting whether a new observation is a "failure" (0) or a "success" (1). The Area Under the ROC curve (AUC) is an aggregated metric that evaluates how well a logistic regression model classifies positive and negative outcomes at all possible cut-offs. AUC is used as a means for evaluating predictive performance of a model although it represents all possible cut-off values. It can range from 0.5 to 1, in our result we have attained a value of *0.85*.

```
final_model = grid.best_estimator_
predict_prob = final_model.predict_proba(X_test)
predict_prob[:5]
y_test
predict_prob[:, 1]
from sklearn.metrics import roc_curve, auc, precision_recall_curve
fpr, tpr, thresholds = roc_curve(y_test, predict_prob[:, 1])
print('AUC: {:.2f}'.format(auc(fpr, tpr)))
import matplotlib.pyplot as plt
plt.figure(figsize = (7, 5))
plt.plot(fpr, tpr, color = 'blue', lw = 2, label = 'Logistic Reg')
plt.plot([0, 1], [0, 1], color = 'red', lw = 2, linestyle = '--', label = 'Random Guess')
plt.title('AUC of Logistic Model')
plt.xlabel('False Positive Rate (1-Specificity)')
plt.ylabel('True Positive Rate' )
plt.xlim([0.01, 1.1])
plt.ylim([0.01, 1.1])
plt.legend()
plt.show()
```

```
AUC: 0.85
```

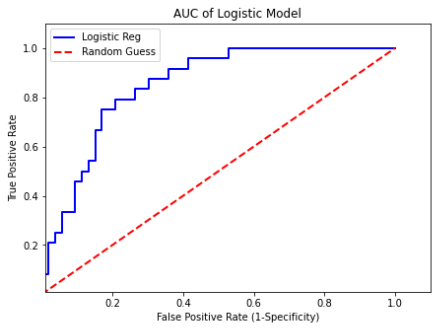


Figure 3.7: AUC of Logistic Model

4. Discussion and Limitations:

By Harshil Chauhan

Working with Data having many Missing Values and Fewer Than 1000 Observations. There are 769 observations and 9 variables in all.

Predictions

1. Blood pressure has a negative influence on the prediction as the higher blood pressure is correlated with a person not being diabetic.
2. Although age was more correlated than BMI to the output variables as concluded during data exploration, the model relies more on BMI. This could happen due to several reasons, including the fact that the correlation captured by age is also captured by some other variable, whereas the information captured by BMI is not captured by other variables.

Logistic Regression

A huge number of categorical features/variables is beyond the scope of logistic regression. Overfitting can be a problem. Furthermore, logistic regression cannot tackle non-linear problems, which is why non-linear features must be transformed. With independent variables that are substantially similar or correlated to each other but not correlated to the target variable, logistic regression will not perform effectively.

Table 2: Summary of what the given data provides as results

Pregnancies	No: 111	Yes: 657				
Glucose	Hyperglycemia: 192	Hypoglycemia: 5	Normal: 571			
BMI	Underweight: 15	Normal: 108	Overweight: 180	Obese: 465		
Diabetes Pedigree Function	Min: 0.0780	1st Qu: 0.2437	Median: 0.3725	Mean: 0.4719	3rd Qu: 0.6262	Max: 2.4200
Age	Min: 21.0	1st Qu: 24.0	Median: 29.0	Mean:33.24	3rd Qu:41.0	Max: 81.0
Outcome	Positive (1): 268	Negative (0): 500				

Type Conversions

Factors

A categorical variable must be created from the outcome variable. We can see that there are about twice as many people who do not have diabetes as there are who do. This should suffice, as there is no uniform limit for the number of rows for the desired variable.

Dealing with Missing Values

6/9 variables in the dataset have several zero markers. It appears that after taking the sum of each column and row separately, there are 763 zero values in the dataset. Alarmingly, this represents almost 100% of our observations.

Numeric

1. **Glucose**: From its existing integer class, it must be changed into numeric variables. These variables contain decimal values, and their absence could lead to erroneous results and skew the risk ranges provided by medical testing. Hyperglycemia was seen in 49 percent of diabetic people, while 50 percent had normal glucose levels. Surprisingly, glucose levels do not appear to distinguish between diabetics and non-diabetics. Obviously, those with hyperglycemia are more prone to develop diabetes, but according to the table above, the risk is relatively low. 87 percent of people without diabetes had normal glucose levels, which is unsurprising.
2. **Pregnancies:** A value of '0' does not always imply that something is missing. For example, a woman's pregnancy record is blank since she has never been pregnant. This is an illustration of the importance of being cautious while preparing data for missing values. When the variable is zero, techniques like regression can provide an estimate of the result. Another option for dealing with zeros is to bin the variable, resulting in a categorical variable. It seems that having a pregnancy does not necessarily increase your chances of having diabetes as the same proportion of women who had or didn't have diabetes had at least one pregnancy.
3. **Insulin:** In Insulin, 50% of the rows have no values. The body generates little or no insulin at times, which is a hallmark of Type 1 Diabetes. Insulin is such a crucial variable in Diabetes, yet when a variable is plagued with missing values, something needs to be done. You could just make up the numbers, but this is medical data, thus half of the numbers are missing. In my

opinion, it would not be fair to just impute the rows with it mean. As a result, no matter how vital it is, it must be removed.

4. **Skin Thickness:** Skin Thickness is in the same boat. The integer 0 appears in 31% of the rows. This variable isn't useful as well.
5. **BMI:** The BMI can be enhanced with its own set of obesity criteria. Because BMI only has 11 0 values, it is unlikely to cause any problems. If it was significantly higher, binning would be ineffective since the allotted bin might not be the correct one.
6. **Obesity:** Unsurprisingly, 80% of Diabetic people were obese while 16% were overweight. Only 3% were reported to be of normal weight. Among the people that do not have diabetes, 50% were obese, 27% overweight and 20% normal.
7. **Diabetes Pedigree Function:** Interestingly, it does not seem to give a clear picture of a diabetic outcome. This is supposed to be a score wherein the higher the score, the more likely the person has diabetes. This is also a variable with a lot of noise.

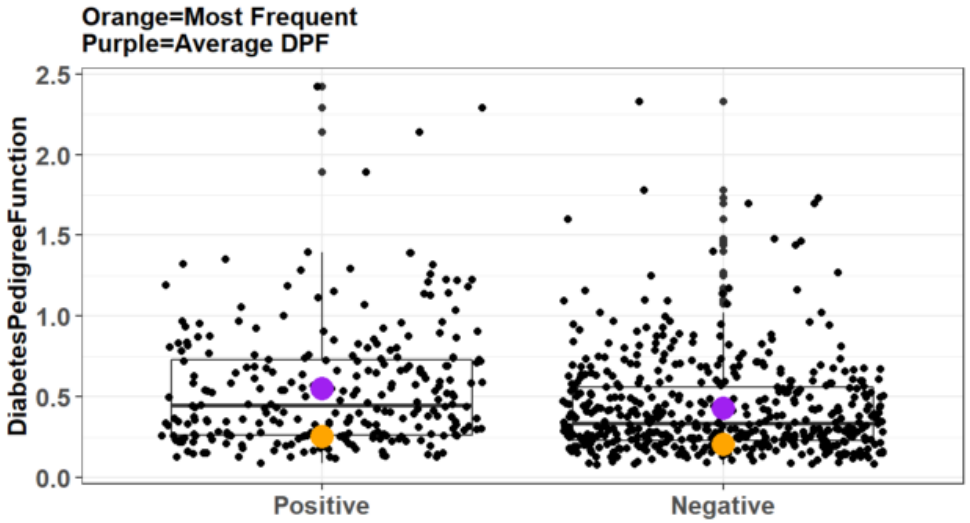


Figure 4.1: Diabetes Pedigree Function by Outcome

5. Conclusion:

By Ramya Datla

With no human intervention required, to predict diabetes one must provide scientific details which includes age, BMI, and so on. The set of rules will offer the effects based on the capabilities extracted and consequently here probabilities of mistake been made are very minimal given that there is no human intervention, and they could similarly continue for treatments or different

tactics should quicker. Proposed Logistic model is to anticipate diabetes that data can be valuable as a model to help foresee diabetes. In this examination, we analyzed the connection between difficulties in diabetic patients and their properties.

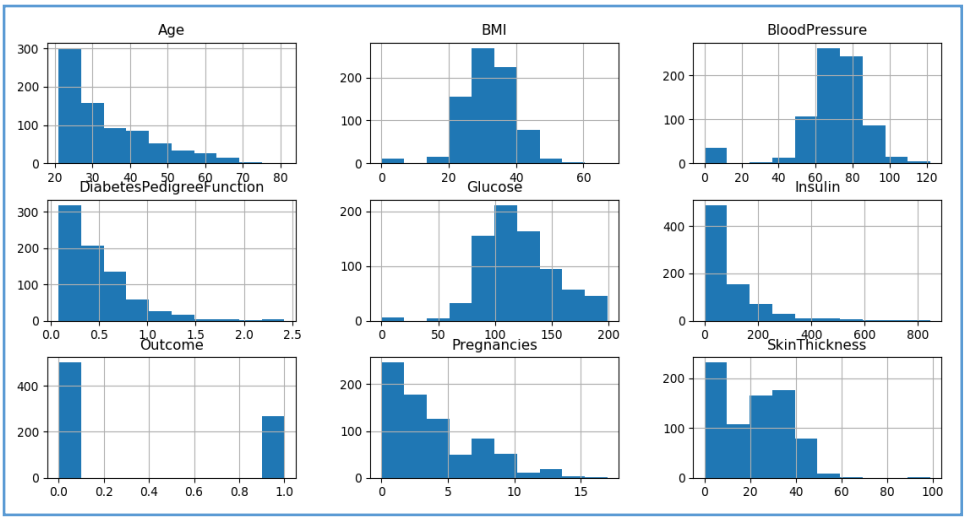


Figure 5.1: Data distribution

The dataset consists of features comprising the medical details of the patients that are useful in determining the health condition of the patient. By using Logistic regression, we import confusion matrix. We compute confusion matrix to evaluate the accuracy of a classification. Accuracy is one metric for evaluating classification models.

Accuracy = Number of CORRECT predictions to Total number of predictions.

The numbers along the diagonal from upper left to lower right represent the correct decisions made, and the numbers outside this diagonal represent the errors. "The true positive rate (also called hit rate or recall) of a classifier is estimated by dividing the correctly classified positives (the true positive count) by the total positive count. The false positive rate (also called a false alarm rate) of the classifier is estimated by dividing the incorrectly classified negatives (the false negative count) by the total negatives. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples.

The Conclusions which can be drawn from the results are:

Glucose level, blood pressure, BMI, pregnancies, skin thickness, diabetes pedigree function have significant and age influence on the model, especially glucose level and BMI. Moreover, the insulins levels play a vital role to determining the model. Interestingly, Our Regression model

match's what we have been hearing from doctors our entire lives, the insulins levels play a critical role in determining the model. Additionally, we notice that blood pressure is also a primary reason as compared to age based on the impact it has in terms of the magnitude. However, BP has a negative influence on the prediction, as the higher BP is correlated with a person being nondiabetic. Although age was more correlated than BMI to the output variables as concluded during data exploration, the model relies more on BMI. This can happen for several reasons, including the fact that the correlation captured by age is also captured by some other variable, whereas the information captured by BMI is not captured by other variables. The proposed Logistic Regression model shows great prediction accuracy in the dataset by measuring the parameters. This underlying dataset for regression training is suitable to predict the professionals without any experience as it offers Reliability, Maintainability Performance, Portability, Scalability, Flexibility Integration. Using Logistic regression, we have achieved 80.2 % accuracy by predicting our data.

6. Future Enhancements:

By Ramya Datla

Project introduction is the first step in building a system. Basically, it will tell what the application or a system is that we are intended to build, what it will look like, brief describe on the proposed project, setting up the project scope, defining project objective, problem statements of the project and the expected outcome. This stage will be used as a reference to ensure system meet the project scope and project objective.

Diabetes is vital health hassle in human society. This paper has summarized techniques for predication of this disease. Logistic Regression Model showed a few promising bring about different area of clinical diagnose with excessive accuracy. It continues to be an open area waiting to get applied in Diabetes predication. Some strategies of have been discussed and implemented for Diabetes predication, alongside pioneer machine getting to know algorithms. An analytical assessment has been completed for locating out best available algorithm for clinical dataset. In future our purpose is to carry ahead the work of temporal scientific dataset, wherein dataset varies with time and retraining of dataset is needed. The proposed system is Diabetes prediction.

We can enhance this system should be efficient to predict the diseases and suggestion of medications using machine learning. Further the system can be extended to N number diseases existing with proper medications. Our proposed System has the benefits which are powerful, flexible, and easy to use. By enhancing increased efficiency of doctor to Improved patient satisfaction. At the same time, reduce the complications with simple, quick, and more accurate results.

References:

- Tiryaki, S., & Aydın, A. (2014). An artificial neural network model for predicting compression strength of heat-treated woods and comparison with a multiple linear regression model. Construction and Building Materials, 62, 102-108.
- Jackson, J. E. (2005). A user's guide to principal components (Vol. 587). John Wiley & Sons.
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. Sociological methods & research, 33(2), 188-229
- Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012, December). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In BMC proceedings (Vol. 6, No. S2, p. S10). BioMed Central.
- Pohlman, J. T., & Leitner, D. W. (2003). A comparison of ordinary least squares and logistic regression.
- https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-institute-diabetes-digestive-kidney-diseases-niddk
- https://www.researchgate.net/publication/286795308_Prediction_of_diabetes_using_logic_regression
- https://realpython.com/logistic-regression-python/
- https://www.sciencedirect.com/science/article/pii/S2666990021000318