DIABETES PREDICTION USING PIMA DIABETES DATASET

Vipin Rai P, Harshabh Mahant, Saurabh Singh

September 14, 2019

1 Data Overview

1. **Number of data points** : 768
2. **Number of features** : 8
3. **Features** :

   Pregnancies : Number of times pregnant

   Glucose : Plasma glucose concentration (2 hours in an oral glucose tolerance test)

   BloodPressure : Diastolic blood pressure (mm Hg)

   SkinThickness : Triceps skin fold thickness (mm)

   Insulin : 2-Hour serum insulin (mu U/ml)

   BMI : Body mass index

   DiabetesPedigreeFunction - Diabetes pedigree function

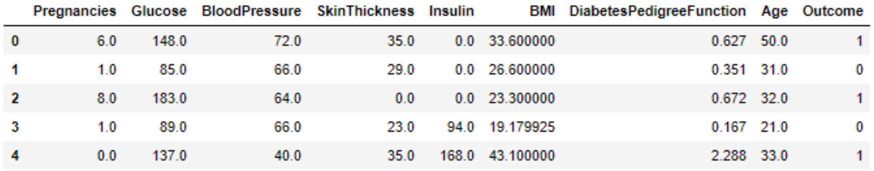
   Age : Age in years
4. **Number of classes** : 2

   class '1' = Patient tested positive for diabetes

   class '0' = Patient tested negative for diabetes
5. **Data points per class** :

   class '1' : 268

   class '0' : 500



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.0	0.0	33.600000	0.627	50.0	1
1	1.0	85.0	66.0	29.0	0.0	26.600000	0.351	31.0	0
2	8.0	183.0	64.0	0.0	0.0	23.300000	0.672	32.0	1
3	1.0	89.0	66.0	23.0	94.0	19.179925	0.167	21.0	0
4	0.0	137.0	40.0	35.0	168.0	43.100000	2.288	33.0	1

Figure 1: Snapshot of dataset

2 Data Preprocessing

2.1 Handling missing values

2.1.1 Observations:

1. Columns containing missing data points cannot be filled with mean values as columns contains values corresponding to different patients. It is absurd to infer the missing data of a patient from distribution of data belonging to other patients.

2. Since the number of datapoints with missing values are less so we can replace the rows with missing values by 0

Feature	**Number of missing values(nan)**
Pregnancies	26
Glucose Value	16
BloodPressure	0
Skin Thickness	22
Insulin	0
BMI	11
Diabetes Pedigree Function	0
Age	19

Table 1: Count of missing values

2.1.2 Inference

1. Features SkinThickness and Insulin contains large number of zero values. Hence these two features can be ignored.

2. Columns Glucose,BloodPressure,BMI,Age will be greater than zero for a patient. Thus rows containing zero value for those features can be eliminated.

3. Pregnancies feature can contain zero values and such rows are retained.

4. In the next step rows containing negative values for features Pregancies,Glucose,BloodPressure, BMI, Age are eliminated. Negative value for those parameters are not valid

5. Rows containing zero values for features Glucose,BloodPressure,BMI, Age are eliminated.

Feature	Number of zero values
Pregnancies	132
Glucose Value	21
BloodPressure	32
Skin Thickness	237
Insulin	374
BMI	21
Diabetes Pedigree Function	0
Age	19

Table 2: Count of zero values after replacing missing values with zeros

3 Exploratory Data Analysis

3.1 Probability Density Function

3.1.1 Observation

1. It can be inferred that patients tested positive for diabetes had higher level of glucose. Most of them had values greater than 130.

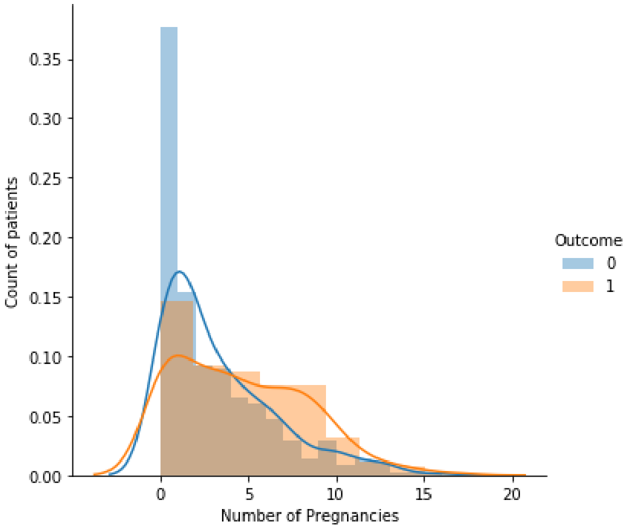
2. BloodPressure values of both the classes patient overlap highly, so it won't be a major feature for classification.

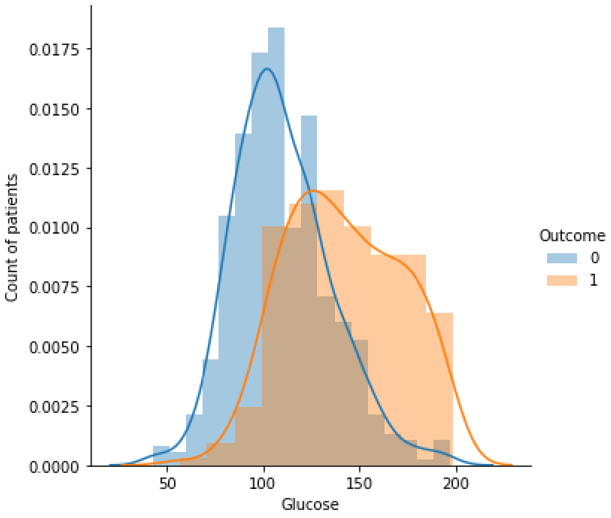
3. Among patient tested positive for diabetes most of them had number of pergnancies greater than 3.

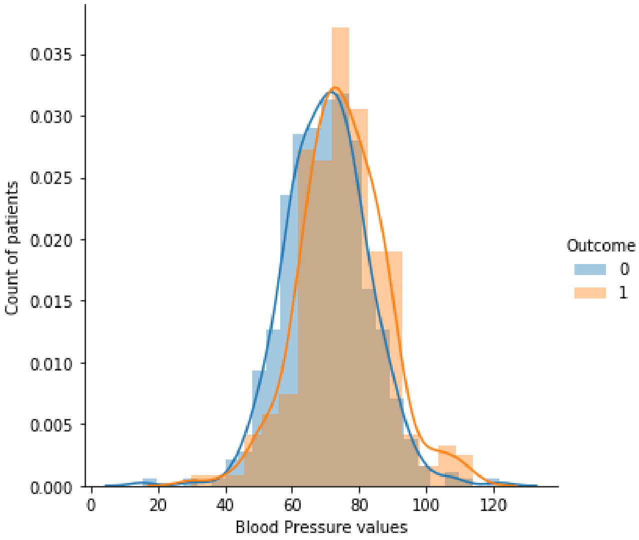
4. Most of the patients with diabetes had BMI value greater than 30.

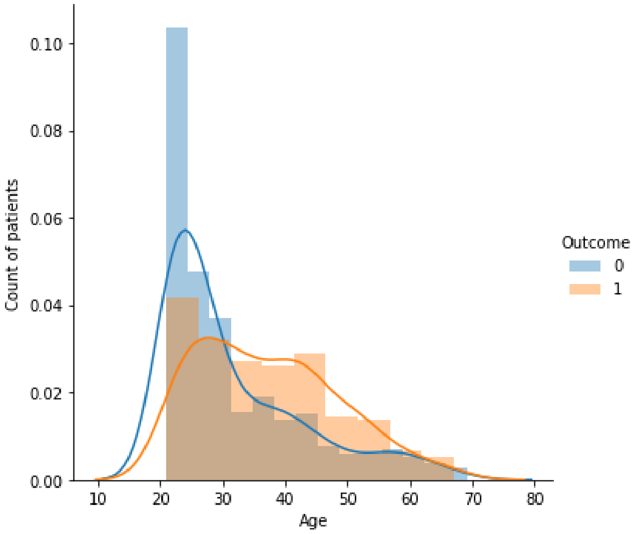
5. Among patient tested negative most of them had diabetes pedigree function value less than 0.5 and if it is greater than 0.5 chances of having diabetes is more.

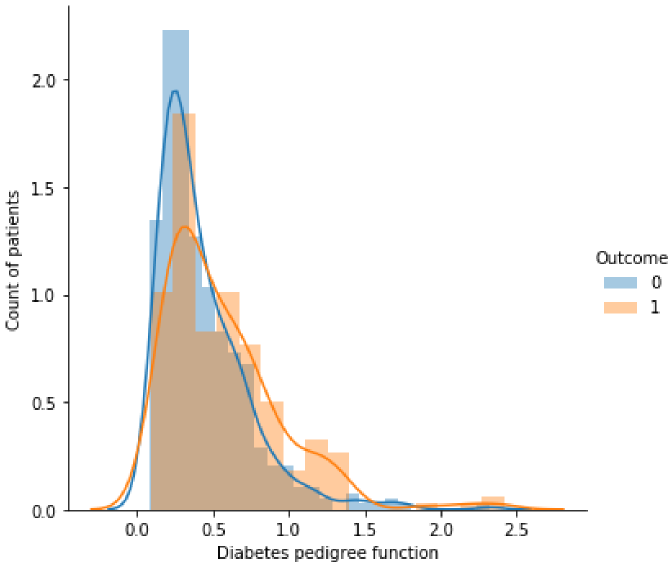
6. Count of patients with diabetes was higher for age group greater than 30.

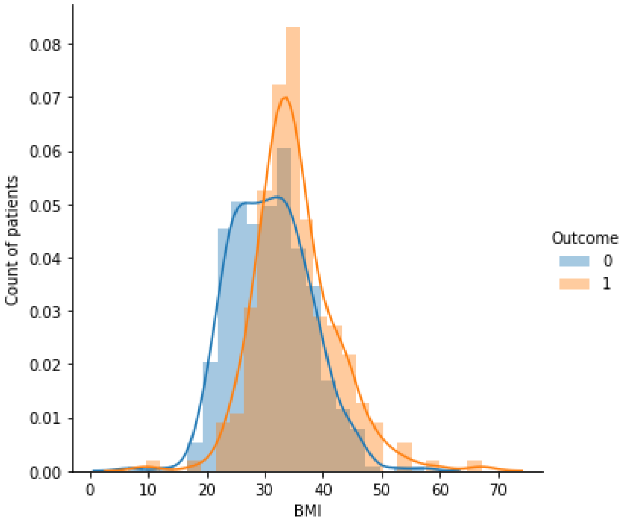

Count of patients
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
Number of Pregnancies
Outcome
0
1


Count of patients
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
50
100
150
200
Glucose
Outcome
0
1


Count of patients
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
Blood Pressure values
Outcome
0
1


Count of patients
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
Age
Outcome
0
1





3.2 Box Plots

1. Box plot can be used to get more insight about the spread of data.
2. It will show the 50th , 25th and 75th percentiles vlaues of a particular feature. The length of the box will give the interquartile range (75th - 25th percentile)

3.2.1 Pregnancies

1. More than 50 percent of the patients tested positive for diabetes had atleast 3 pregnancies
2. 50 percent of patient tested negative had less than 2 pregnancies
3. It can be inferred that patient with more number of pregancies have higher chances of having diabetes

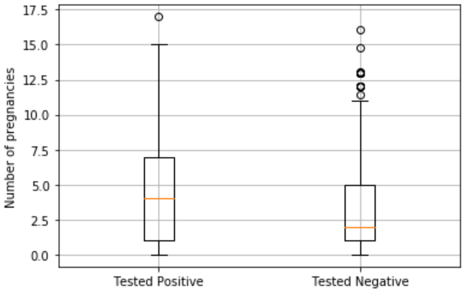


Figure 5: Boxplot for Pregnancies

3.2.2 BMI

1. More than 75 percent of patients with diabetes had BMI greater than 30
2. 50 percent of patients without diabetes had BMI less than 30
3. BMI can be a major feature to decide whethere a patient is diabetic
4. Patient with BMI greater than 30 has higher chances of having diabetes

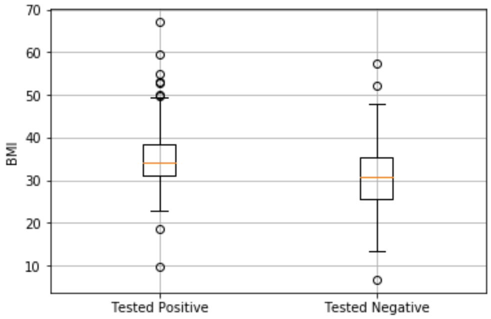


Figure 6: Boxplot for BMI

3.2.3 Glucose

1. More than 70 percent of patients with diabetes had glucode level higher than 120
2. Among patients tested negative 70 percent of them had glucose level less than 120
3. It is highly likely that patient has diabetes if his glucose level is more than 120

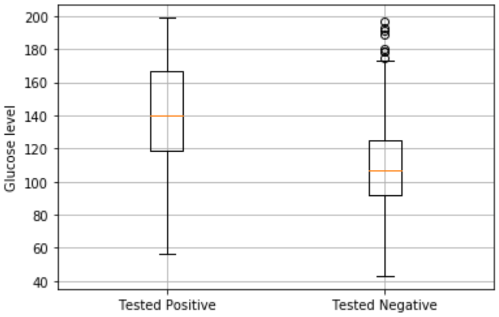


Figure 7: Boxplot for Glucose

3.2.4 Blood Pressure

1. Among of the patient tested negative for diabetes more than 75 percent of them had blood pressure in the range of 60 to 80 mm Hg
2. More than 50 percent of patients with diabetes had blood pressure higher than 70 mm Hg
3. Blood Pressure ranges are mostly overlaping for both classes of patients , however patient with blood pressure greater than 80 mm Hg is likely to have diabetes

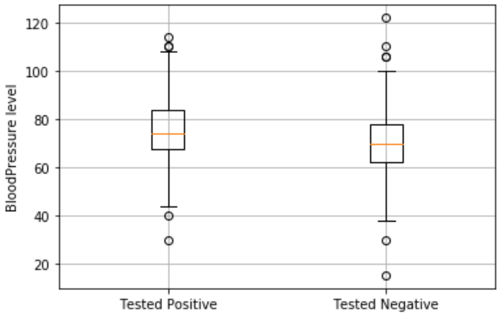


Figure 8: Boxplot for Blood Pressure

3.2.5 Age

1. More than 75 percent of patients tested negative had age less than 38
2. More than 50 percent of patients with diabetes were of age greater than 35
3. Patient with age greater than 35 years have chances of having diabetes
4. Age can be considered as one of the major feature for classification

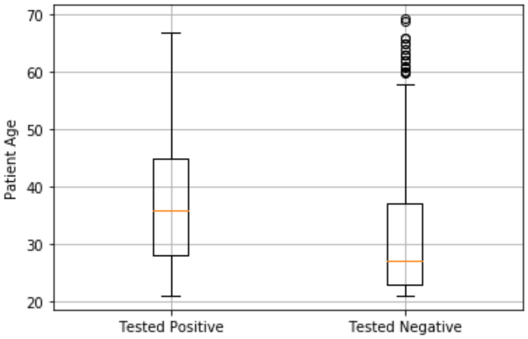


Figure 9: Boxplot for Age

3.2.6 Diabetes Pedigree Function

1. More than 50 percent of patient with Pedigree function value greater than 0.4 had diabetes

2. Among the patients tested negative for diabetes 50 percent of them had diabetes pedigree function vaue less than 0.3

3. If the diabetes pedigree function value is greater than 0.5 it is highly likely to have diabetes

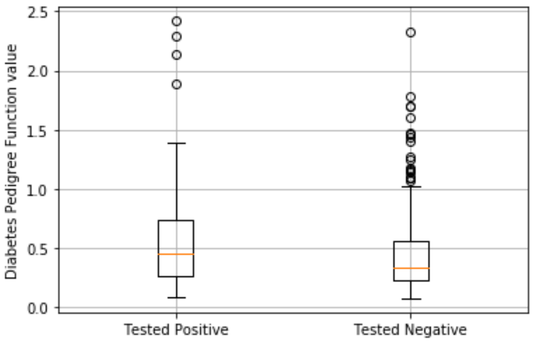


Figure 10: Boxplot for Diabetes Pedigree Function

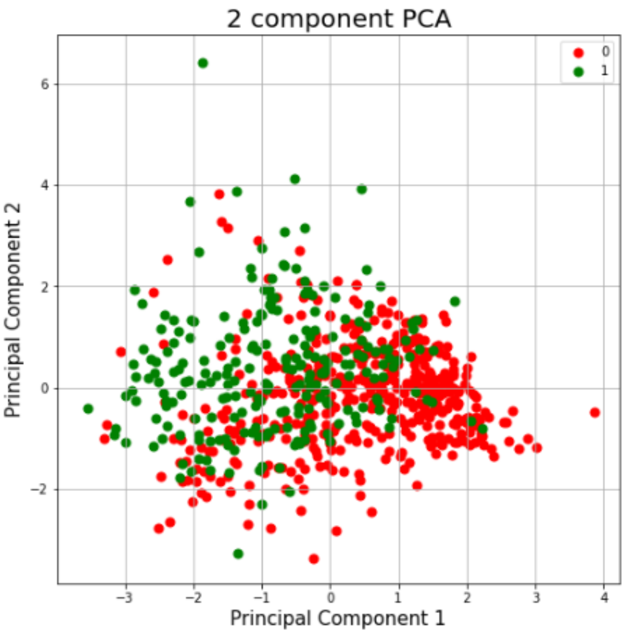
4 Model Building

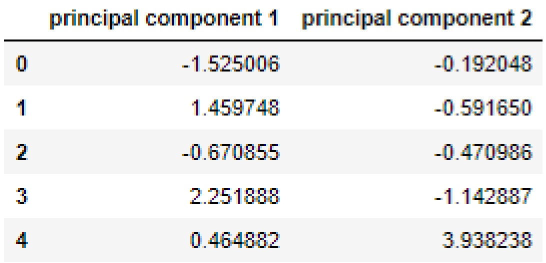
1. Features taken into cosideration ["Pregnancies","Glucose","BloodPressure", "BMI", "Diabetes Pedigree Function", "Age"]

2. Since it's a classification problem logistic regression and knn algorithms are used to train the model.

4.1 Dimensionality Reduction using PCA

1. PCA can be used to reduce the dimension from 6 features to 2 main features. Accuracy of model will be evaluated with and without dimensionality reduction.





	principal component 1	principal component 2
0	-1.525006	-0.192048
1	1.459748	-0.591650
2	-0.670855	-0.470986
3	2.251888	-1.142887
4	0.464882	3.938238

4.2 Spliting dataset into train and test

1. The given data is split into train and test in the ratio 8:2.

4.3 Using Logistic Regression

4.3.1 Logistic regression without dimensionality reduction(without PCA)

1. Accuracy of Logistic Regression without pca : 80.14705882352942

	Class 0	Class 1
Class 0	86	4
Class 1	23	23

Table 3: Confusion matrix

4.3.2 Logistic regression with dimensionality reduction(PCA)

1. Accuracy of Logistic Regression with pca : 75.0

	Class 0	Class 1
Class 0	86	4
Class 1	30	16

Table 4: Confusion matrix

4.4 Using K-Nearest Neighbour

1. From the scatter plot(Principal component 1 vs principal component 2) it can be inferred that two classes are not linearly separable.

2. Logistic regression might be failing to fit a accurate plane that can separate the two classes with more accuracy.

3. K-NN algorithm finds k-nearest neigbours for a data point and decides it's class based on majotity rule.

4. Best values of k can be chosen by a technique called cross validation (10 fold). Here the train data is split into 10 parts. In each interval one part among 10 is chosen as test and knn is applied on that. Mean of accuracy is calculated for all 10 parts(cv scores). This is repeated for different values of k. k value that gives best accuracy is chosen as optimal k value.

5. Model is designed for dataset with and without pca and accuracy is tested for both cases.

4.4.1 K-NN without dimensionality reduction(without PCA)

1. The optimal value of k(neighbours) = 33.

2. Accuracy of knn without pca : 83.088235

	Class 0	Class 1
Class 0	88	2
Class 1	21	25

Table 5: Confusion matrix

4.4.2 K-NN with dimensionality reduction(PCA)

1. The optimal value of k(neighbours) = 41.

2. Accuracy of knn with pca : 75.73529411764706

	Class 0	Class 1
Class 0	86	4
Class 1	29	17

Table 6: Confusion matrix

5 Conclusion

1. Data preprocessing was done to replace the missing values with appropriate ones. The rows containing invalid inputs(like zero values) for certain columns were eliminated.

2. Exploratoty data analysis was made to get insight about the distribution of the data with respect to different features. Probability density function and Box plot tools were used for analysis.

3. PCA was used for dimensionality reduction and principal components were plotted using 2d-scatter plot.

4. Model was trained using Logistic regression. Model was trained with and without dimensionality reduction.Accuracy was ranging between 75 to 80 percent.

5. Model was trained using K-NN. Optimal value for k was chosen using cross validation technique. Model was trained with and without dimensionality reduction.Accuracy was ranging between 75 to 80 percent.