DIABETES PREDICTION WITH ML

Sabin Maharjan BS Computer Science, TU

Abstract

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have a high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospitals is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays a significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In the existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with the new dataset compared to the existing dataset. Further imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

Keywords: Diabetes Prediction, Diabetes Mellitus, Logistic Regression, SVM, KNN, Diabetes Database, Predictive Analysis, etc.

1. Introduction

Diabetes mellitus (DM), commonly referred to as diabetes, is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Type 1 diabetes results from the pancreas's failure to produce enough insulin. Type 2 diabetes begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As of 2015, an estimated 415 million people had diabetes worldwide, with type 2 diabetes making up about 90% of the cases. This represents 8.3% of the

adult population.mHealthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyzes huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data.

Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are suffering from it.
Diabetes Mellitus (DM) is classified as-

Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient.

Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.

Type-3 Gestational Diabetes, increase in blood sugar level in pregnant women where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person.

1.1 Problem Definition

A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression techniques. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes. Machine learning is considered to be one of the most important artificial intelligence features that supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing methods for diabetes detection use lab tests such as fasting blood glucose and oral

glucose tolerance. However, this method is time consuming. This paper focuses on building predictive models using machine learning algorithms and data mining techniques for diabetes prediction.

1.2 Objective

- The objectives of the proposed project can be listed as follows:
- To develop a model to predict diabetic patients out of a given dataset.
- To find an accurate algorithm for the prediction.
- To study and analyze the percentage of diabetic patients and different factors affecting it.
- To analyze the impact of each factor
- To model and evaluate approaches
- To balance the dataset

2. Methods

The design methodology includes:

- Dataset Collection
- Data Pre-processing
- Clustering
- Model Building

2.1 Dataset Collection

This module includes data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results. Dataset description is given below-

This Diabetes dataset contains 800 records and 10 attributes.

Table 1. Dataset Information

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness (mm)	N
Insulin	N
BMI	N
Age	N
DiabetesPedigreeFunction	N
Outcome	C

2.2 Data Pre-processing

This phase of the model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for a few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

Data preprocessing is the most important process. Most healthcare related data contains missing value and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after the mining process, Data

preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps

2.2.1 Missing Values removal

Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and helps to work faster.

2.2.2 Splitting of data

After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithms on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically the aim of normalization is to bring all the attributes under the same scale.

2.3 Clustering

In this phase, we have implemented K-means clustering on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found which were, Glucose and Age. K-means clustering was performed on these two attributes. After implementation of this clustering we got class labels (0 or 1) for each of our records.

Algorithm for Clustering:

- Choose the number of clusters(K) and obtain the data points
- Place the centroids c_1, c_2, c_k randomly
- Steps 4 and 5 should be repeated until the end of a fixed number of iterations
- For each data point x_i:
- Find the nearest centroid(c_1, c_2 ,..c_k)
- Assign the point to that cluster
- for each cluster j = 1..k
- new centroid = mean of all points assigned to that cluster
- End

2.4 Model Building

Algorithm: Diabetes Prediction using various machine learning algorithms

- Generate training set and test set
- randomly. Specify algorithms that are
- used in model mn=[KNN(), SVC(), RandomForestClassifier(), LogisticRegression()]
- for(i=0; i<13; i++) do
- Model= mn[i];
- Model.fit();
- model.predict();
- print(Accuracy(i),confusion_matrix,
- classification_report);
- End

2.4.1 Support Vector Machine

Support Vector Machine also known as SVM is a supervised machine learning algorithm. Svm is the most popular classification technique. Svm creates a hyperplane that separates two classes. It can create a hyperplane or set of hyperplanes in high dimensional space. This hyperplane can be used for classification or regression also. Svm differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by a hyperplane that performs the separation to the closest training point of any class.

Algorithm

- Select the hyper plane which divides the class better.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
- Select the class which has the highest margin. Margin = distance to positive point + Distance to negative point.

2.4.2 K-Nearest Neighbor

KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is a lazy prediction technique.KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other.KNN helps to group new work based on similarity measure.KNN algorithm records all the records and classifies them according to their similarity measure. For finding the distance between the points uses a tree-like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — its nearest neighbors. Here K= Number of nearby neighbors, it's always a positive integer. Neighbor's value is chosen from the set of classes. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, ... Pn) and Q (q1, q2,..qn) is defined by the following equation:-

Algorithm

- Take a sample dataset of columns and rows.
- Take a test dataset of attributes and rows. · Find the Euclidean distance by the help of formula-

$$EculideanDistance = \sqrt{\sum_{i=1}^{y}\sum_{j=1}^{m}\sum_{l=1}^{n-1}(R_{(j,l)} - P_{(i,l)})^2}$$

- Decide a random value of K. is the no. of nearest neighbors
- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values. If the values are the same, then the patient is diabetic, otherwise not.

2.4.3 Logistic Regression

Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories. It classifies the data in binary form only

in 0 and 1 which refer to a case to classify a patient that is positive or negative for diabetes. Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variables. Logistic regression is based on a Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

Sigmoid function P = 1/1+e - (a+bx)
Here P = probability, a and b = parameter of Model.

2.4.4 Random Forest

It is a type of ensemble learning meth- od and also used for classification and regression tasks. The accuracy it gives is greater than compared to other models. This method can easily handle large datasets. Ran- dom Forest is developed by Leo Bremen. It is a popular en- semble Learning Method. Random Forest improves Performance of Decision Trees by Reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the indi- vidual trees.

Algorithm

- The first step is to select the “R” features from the total features “m” where R<<M.
- Among the “R” features, the node uses the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until the ”l” number of nodes has been reached.
- Built forest by repeating steps a to d for “a” num- ber of times to create “n” number of trees.
- The random forest finds the best split using the Gin-Index Cost Function which is given by:

$$Gini = \sum_{k=1}^{n} p_k * (1 - p_k) \; Where \; k = Each \; class \; and \; p = proption \; of \; training \; instances$$

The first step is to take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and store the anticipated outcome at intervals at the target place. Secondly, calculate the votes for each

predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest do correct predictions result for a spread of applications.

2.4.5 Naïve Bayes

The naïve Bayes is the simple supervised machine learning algorithm based on the Bayes theorem. The algorithm assumes that the features conditions are independent of the given class. The naïve Bayes algorithm helps to build fast machine learning models that can make a fast prediction. The algorithm finds whether a particular portion has a spot by a particular class and utilizes the probability of likelihood.

2.4.6 Gradient Boosting

Gradient Boosting is the most powerful ensemble technique used for prediction and it is a classification technique. It combines weak learners together to make strong learner models for prediction. It uses the Decision Tree model. It classifies complex data sets and it is a very effective and popular method. In gradient boosting model performance improves over iterations.

Algorithm

- Consider a sample of target values as P
- Estimate the error in target values.
- Update and adjust the weights to reduce error M. P[x] =p[x] +alpha M[x]
- Model Learners are analyzed and calculated by loss function F
- Repeat steps till desired & target result P.

3. Testing

This is the final step of the prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score.

3.1 Classification Accuracy

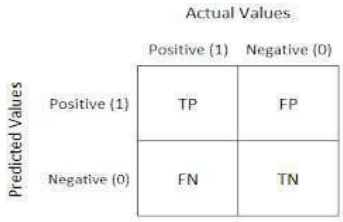
It is the ratio of the number of correct predictions to the total number of input samples. It is given as

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

3.2 Confusion Matrix

It gives us a matrix as output and describes the complete performance of the model.

Where, TP: True Positive
FP: False Positive FN:False
Negative TN: True Negative



3.3 F1 score

It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

$$F1 = 2 * \frac{1}{\left(\frac{1}{precision}\right) + \left(\frac{1}{recall}\right)}$$

3.4 Precision

It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as-

$$Precision = \frac{TP}{(TP + FP)}$$

3.5 Recall

It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

$$Precision = \frac{TP}{(TP + FP)}$$

3.6 ROC AUC Curve

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

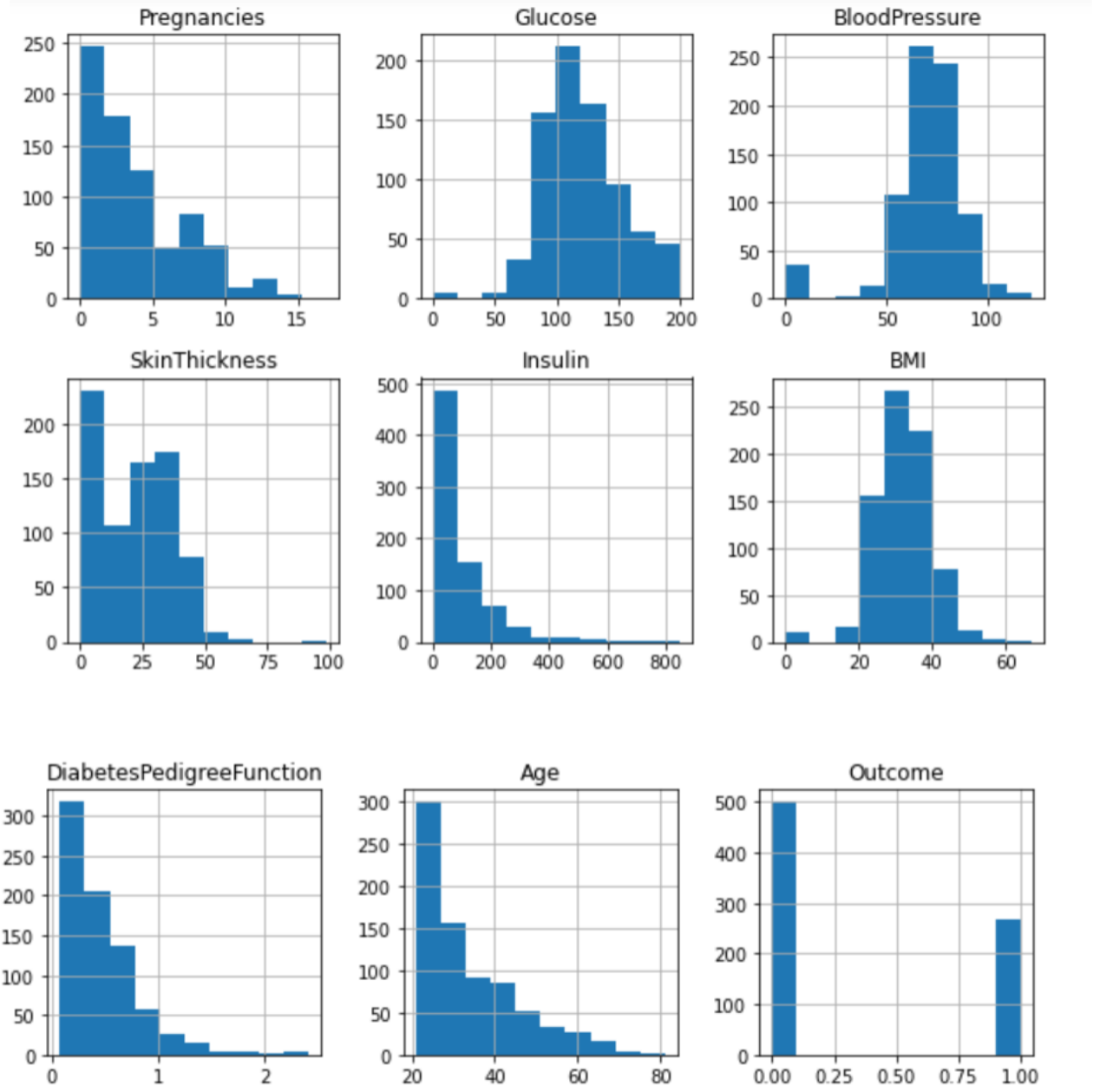
4. Results

After applying various Machine Learning Algorithms on the dataset we got accuracies as mentioned below. Random gives the highest accuracy of 84.375% and ROC of 80.12%.

4.1 DATASET

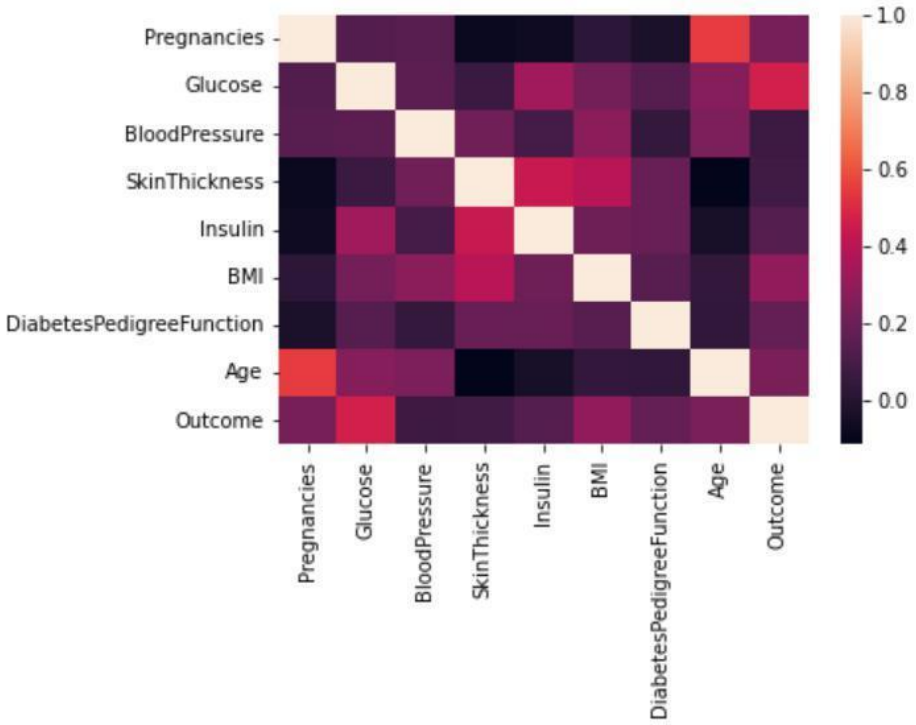
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	138	62	35	0	33.6	0.127	47	1
0	84	82	31	125	38.2	0.233	23	0
0	145	0	0	0	44.2	0.63	31	1
0	135	68	42	250	42.3	0.365	24	1
1	139	62	41	480	40.7	0.536	21	0
0	173	78	32	265	46.5	1.159	58	0
4	99	72	17	0	25.6	0.294	28	0
8	194	80	0	0	26.1	0.551	67	0
2	83	65	28	66	36.8	0.629	24	0
2	89	90	30	0	33.5	0.292	42	0
4	99	68	38	0	32.8	0.145	33	0
4	125	70	18	122	28.9	1.144	45	1
3	80	0	0	0	0	0.174	22	0
6	166	74	0	0	26.6	0.304	66	0
5	110	68	0	0	26	0.292	30	0
2	81	72	15	76	30.1	0.547	25	0
7	195	70	33	145	25.1	0.163	55	1

4.2 HISTOGRAM



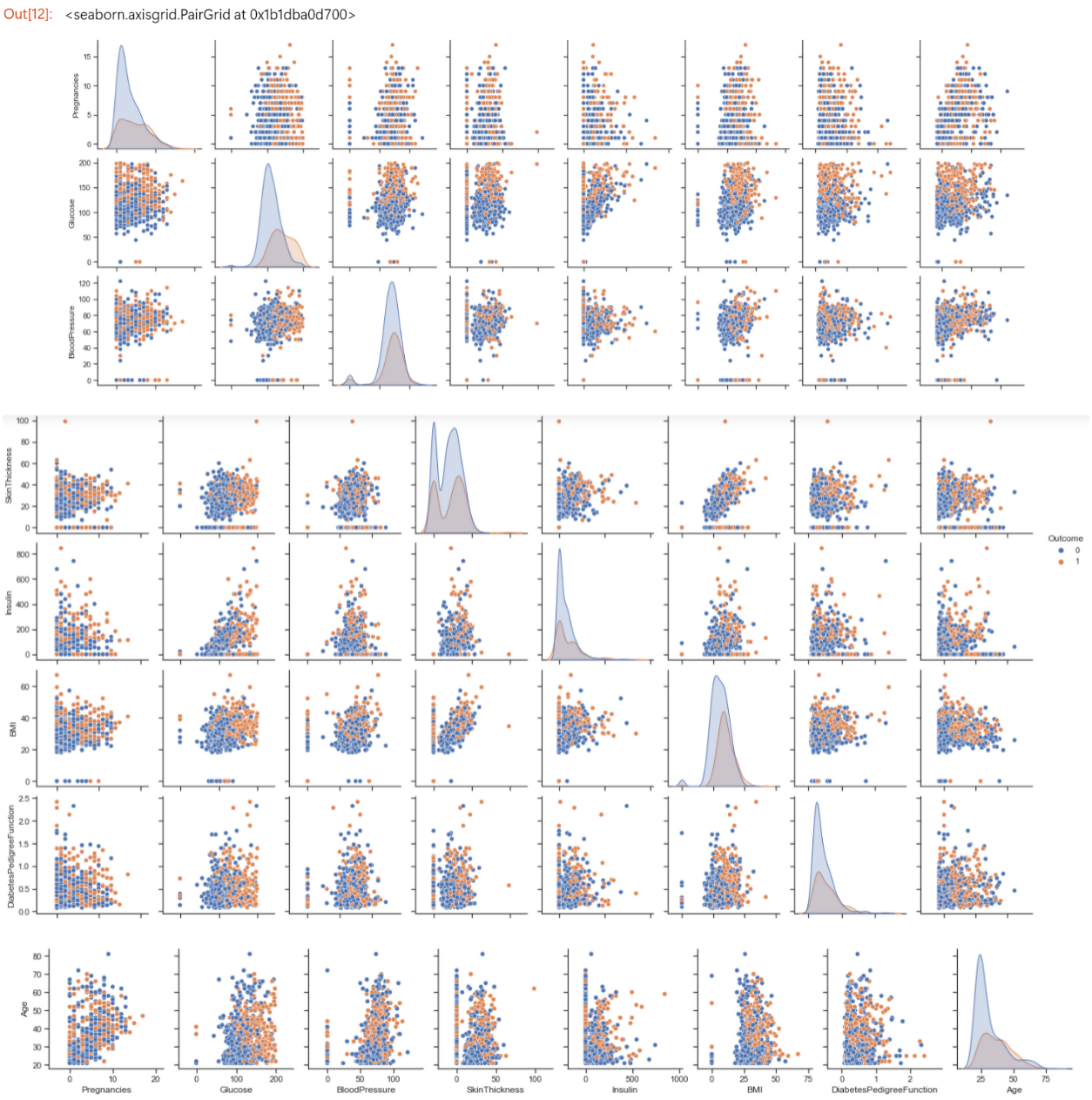
4.3 CORRELATION

Out[10]: <AxesSubplot:>

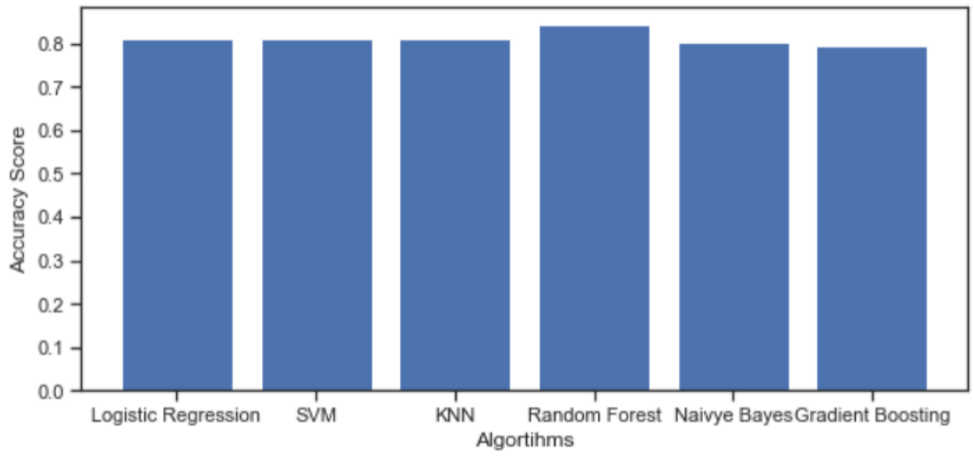


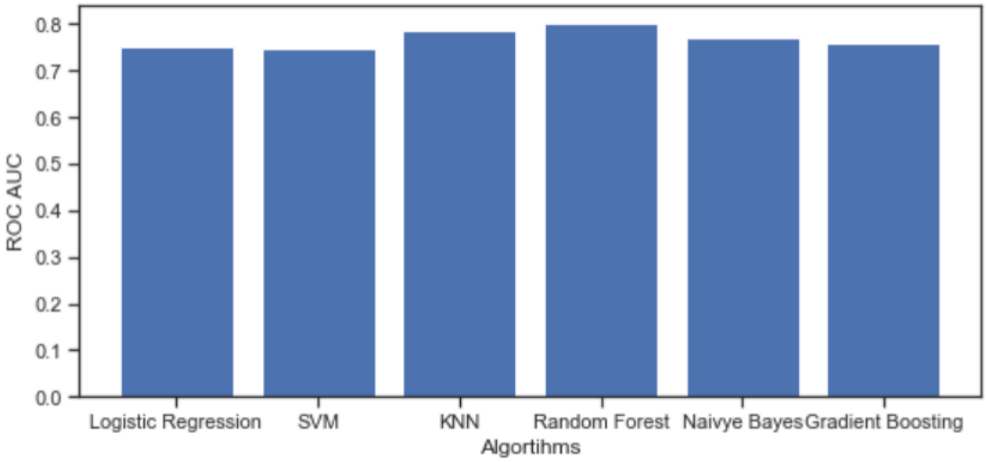
4.4 PLOT

Out[12]: <seaborn.axisgrid.PairGrid at 0x1b1dba0d700>



4.5 BAR GRAPH FOR ACCURACY AND ROC SCORE





References

T. M. Alama, M. A. Iqbal, Y. Ali et al., “A Model for Early Prediction of Diabetes,” Informatics in Medicine Unlocked, vol. 16, Article ID 100204, 2019.

M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare,” in Proceedings of the 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, UK, September 2018.

Md. Maniruzzaman, Md. Jahanur Rahman, B. Ahammed, and Md. Menhazul Abedin, “Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm,” Health Information Science and Systems, vol. 8, 2020.

Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I

B. Nithya and Dr. V. Ilango,” Predictive Analytics in Healthcare Using Machine Learning Tools and Techniques”, International Conference on Intelligent Computing and Control Systems

Diabetes Dataset by PIMA INDIANS