JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

Machine Learning for Smartphone-Based Early Detection of Diabetic Disease in Pima Indians Diabetes Database

[1]Subash Chandra Bose Jaganathan, [2]Chandramohan Dhasarathan, [3]Ivasapuram Harshavardhan,[4]Chinnam Harisrujan, [5]Gogineni Haridhar, [6]Gude Ganga Prasad, S.Kannadhasan[7]
[1]School of Computing Science & Engineering,
VIT Bhopal University, MP, India
[2]Computer Science & Engineering Department,
Thapar Institute of Engineering & Technology, Patiala, Punjab, India.
[3,4,5,6]Department of Computer Science & Engineering,
[7] Department of Electronics and Communication Engineering, Cheran College of Engineering, Karur, Tamilnadu, India
Madanapalle Institute of Technology & Science,Madanapalle, Andra Pradesh, INDIA.
jsubashme@gmail.com, pdchandramohan@gmail.com, kannadhasan.ece@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10.

Abstract:
Now a day's Diabetic disease(Dd) is common disease among people which causes damage to kidney, heart and may eventually lead to death. Early detection of diabetics is very much essential to avoid kidney and heart failure. Effective treatment for Dd are available through it requires early diagnosis and the continuous monitoring of diabetic patients. Also many physical tests can be used to detect Dd but are time consuming. The objective of our research paper is to give decision about the presence of diabetics by applying ensemble of machine learning classifying algorithms on features extracted from output of different datasets. It will give us accuracy of which algorithm will be suitable and more accurate for prediction of the disease. Decision making for predicting the presence of diabetic is performed using Linear Regression(LR), Logistic Regression(LoR), K Nearest Neighbors(KNN), Decision Tree(DT), Support Vector Machine(SVM), Naïve Bayes(NB), Random Forest(RF), The experimental results has been analysed by using Jupyter Notebook. Among all the mentioned Supervised machine learning algorithm RF approach show a highest classification accuracy (CA) of 89.58. From this, we can infer that for diabetic the RF approach gives the best performance compared to all other approaches.
Keywords*:* Supervised Learning algorithms, Machine Learning, Diabetes, Classification accuracy, precision, recall, f1-score,

Introduction:

Diabetics are achronic and organ disease that occurs when the pancreas does not secrete enough insulin or the body is unable to process it properly. Overtime, diabetics affect the circular system. Diabetic is a medical condition where it damaged the human organs because of fluid leaks from blood vessels into the human body. According to World health organization (WHO) 415 million diabetics

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

patients are at risk of losing their life because of diabetics. It occurs LR, LoR, KNN, DT, SVM, NB, RF. Can be trained by providing training datasets to them and then these algorithms can predict the data by comparing the provided data with the training datasets. Our objective is to train our algorithm by providing training dataset to it and our goal is to detect diabetic using different types of classification algorithms.

Machine Learning(ML) a branch of Artificial Intelligence(AI), concerns, the constructions and study of systems that can learn from data [4]. ML algorithms use computational methods to learn information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance a widely quoted and more formal definition:

A computer program is said to learn from experience E with respect to same class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves with experience E [5].

The core of machine learning deals with representation and generalization. Representing the data instances and functions evaluated on these instances are part of all machine learning systems.

Generalizations is the ability of a machine learning system to perform accurately on new, unseen data instances after having experienced a leaning data instance. The training examples come from some generally unknown probability distribution and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases. The performance of generalization is 11 usually evaluated with respect to the ability to reproduce known knowledge from newer examples. There are different types of machine learning but the two main ones are Supervised learning and Unsupervised learning.

Supervised Learning Model

Supervised learning model is the machine learning task of inferring a function from supervised training data [6].

Training data for supervised learning includes a set of examples with paired input subjects and desired output. A supervised learning algorithm analyses the training data and produces an inferred function, which is called classifier or a regression function. The function should predict the correct output value for the any valid input object. This requires the learning algorithm to generalize from the training data to unseen situation in a reasonable way.

A simple analogy to supervised learning is the relationship between a student and a teacher. Initially the teacher teaches the student about a particular topic. Teaching the students the concepts of the topic and then giving answers to many questions regarding the topic. Then the teacher sets as exam paper for the students to take, where the students answers the newer questions.

Figure1 describes that the system learns from the data provided which contains the features and the outputs as well. After it has done learning, newer data is provided without outputs, and the system generates the output using the knowledge it gained from the data on which it trained. Here is how supervised learning model works.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

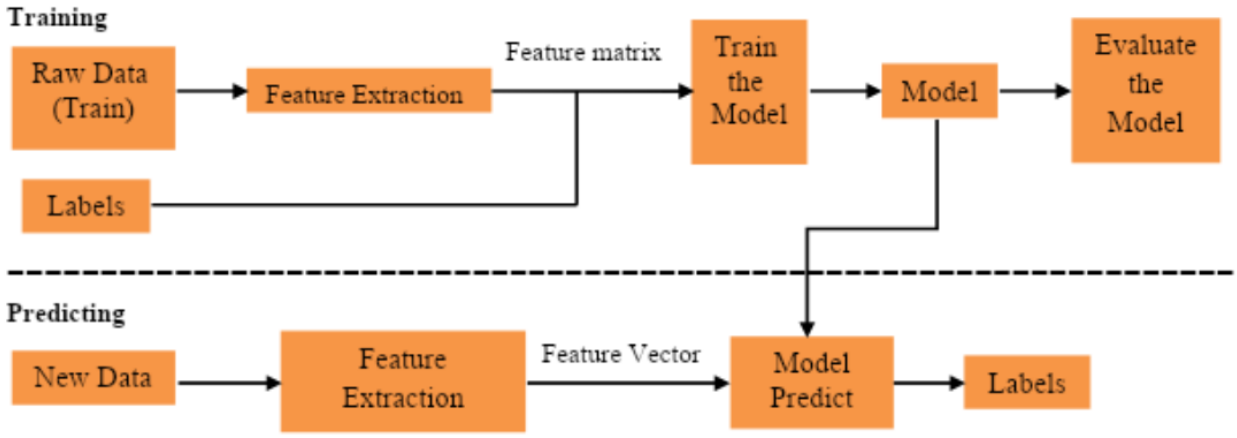

Figure 1. Flow diagram

Dataset:

Pima Indian Diabetes dataset contains 768 patients details with 8 attributes the sample data appearance has been shown in the Figure 2.

Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
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	1
116	74	0	0	25.6	0.201	30	0
78	50	32	88	31	0.248	26	1
115	0	0	0	35.3	0.134	29	0
197	70	45	543	30.5	0.158	53	1

Figure 2. Sample input data

Algorithms:

Since there are so many algorithms for ML, it is not possible to use all of them for analysis. For this research paper, we will be using: LR, LoR, KNN, DT, SVM, NB, RF

RF:

RF algorithm can use both for classification and the regression kind of problems. It is supervised classification algorithm which creates the forest with a number of tress [9]. In general, the more trees in the forest the more robust the forest looks like. It could be also said that the higher the number of trees in the forest gives the high accuracy results. There are many advantages of random forest algorithms. The classifier can handle the missing values. It can also model the random forest classifier for categorical values [10]. The over fitting problem will never come when we use the random forest algorithm in any classification problem. Most importantly it can be used for feature engineering which means identifying the most important feature out of the available feature from the training dataset.

K - Nearest Neighbors

K - Nearest Neighbors is a simple algorithm that score all available cases and classifies new cases based on a similarity measure [11]. KNN has been used in statistical estimation and pattern recognition. KNN makes prediction for new instance(x) by searching

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

through the entire training set for the k most similar instances and summarizing the output variable for those k instances. For regression this might be the means output variable, in classification this might be the mode class determine which of the k instances in the training dataset are most similar to new input many distance measure is used like Euclidean distance, Manhattan distance, Minkowski distance.

Distance Functions

Euclidean = $\sqrt{\sum_{i=1}^{k}(Xi - Yi)2}$ $(x + a)^n = \sum_{k=0}^{n} x^k a^{n-k} \binom{n}{k}$

Manhattan = $(x + a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$

Logistic Regression

In statistics, the logistic model (or logit model) is a widely used statistical model. In its basic form it uses a logistic function to model a binary dependent variable although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model(a form of binominal regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/risk these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, I=the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors") the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certain the value "0") and 1 (certainly the value "1"), hence the labeling the function that converts log-odds to probability is the logistic function, hence the name [12].

Literature Survey

Alshamlan et. al. predicted diabetic and normal persons by using fisher score feature selection, chi-2 feature selection and Logistic Regression supervised learning algorithm with best accuracy of 90.23%. Mainly they train the model based on the 212 million data elements that is collected through out the world. Their development stages as follows loading the data set ,filtering the method and classifier[13].Usman et. al. predicted diabetic and non-diabetic people by processing two datasets. Feature selection with logistic regression classification were used. The obtained accuracy result of logistic regression on two datasets based on fisher score feature selection was higher than Ch-2 feature selection. The accuracy results of two data were 90.23% and 61.90% respectively. In this paper they mainly predict the rick of having increased Artificial Stiffness among diabetic Patients using logistic regression. They mainly take the value of auc-ppg, age and HbA1-c values as input and predict the rick level using statistical model (Logistic regression).They take data set from the kpp2 clinic for train the model. They set the values as high risk is form 32-37 and low risk as below this values. Overall they get the accuracy almost 93% for the trained modle[14].*Chaudhuri et. al.* gathered information from the Pima Indians Diabetes Database(www.kaggle.com). They mainly consider three data sets to predict the diabetes. Diabetes disease dataset, Liver disease dataset and Heart Disease(HD) dataset are the data sets they are considering. The prediction accuracy for their proposed model is 97%. This paper applied GA to

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

predict the progression of the disease by integrating the effects of a large set of independent variables from datasets (n number of possible combinations within the minimum and maximum values) with the best LR formula chosen to calculate the best accuracy. The classification result of their proposed classifier shows that the combination of variables is superior in quality to the use of a single significant variable or a finite set of variables to predict disease progression[15].Ahmed et. al. developed an applicable system to predict diabetes using distributed machine learning based on big data platforms such as Spark. In this context, their study aims to develop models using distributed machine learning based on Apache Spark to predict diabetes. Five machine learning classification methods were used like Decision Tree, Support Vector Machine, Logistic Regression Classifier, Naive Bayes, and Random Forest Classifier. Comparison between different algorithms was calculated using three measures, which are accuracy, recall, and precision. The experimental results proposed that LR achieved the highest percentage of accuracy, recall, and precision,82%, 92%, and 82%, respectively. So they predicted the diabetes based on the Lenear regression model. Accuracy is more than 82% for their trained model[16].*Le* et. al. actually using heat map to calculate correlation between the features . After understanding the correlation between the different features they apply the T-Test to extract the features from the actual data set. After testing with the different machine learning algorithms they come with an most accurate machine learning algorithm that is Logistic Regression. Using this methos they are getting more than 79% accuracy. After all testing's they decided to move on with the Logistic regression algorithm. The implemented the logistic regression with basic Relu formula and analyses its behavior on the input data[17].Barhate et. al.they take the data from the Pima all India diabetes data set. They are actually divide the data set into two parts 70% and 30% respectively. The high data set part is for training and low data set part is for testing. They are actually using 7 different machine learning algorithms to predict the diabetes. Finally random forest gives the highest accuracy on the Pima data set . It is giving about 79.9% accuracy against test data set. So they concluded best algorithm is Random forest machine learning algorithm for predicting the diabetes[18].Anirudh Hebbar et. al. mentioned about predicting diabetes. Predicting the unknown substance using the existed value. The predicted value is only the two values , that are either 0(No) or 1(Yes). To predict this diabetes using existed values they used machine learning algorithms such as decision tree and random forest. Features considered for the constructing Decision Tree and Random Forest Based Classification Model to Predict Diabetes are glucose level, blood pressure, insulin, body mass index, age, etc. The machine learning algorithm is evaluated on real-life data set. Finally they got accuracy of 72% for the Decision tree and 76.5% for Random Forest[19].VijiyaKumar et. al.predicted the diabetes using Random forest Algorithm. They taken the real time diabetic patient data set. First they take the new input value and they pre process that data. Pre processing involves many technique's like data cleaning ,data reduction and data integration. Than here the patient real time data set may be imported . After that they implement the Random Forest algorithm using that data set . Training the model and test that model involve in this process only[20] In this

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

implementation process they implement the different trees from the data set and Sivaranjani et. al. predicted the diabetes using the Machine Learning algorithms Support Vector Machine (SVM) & Random Forest (RF) . First they take the data set from a clinic and preprocess the data set . In the preprocessing of the data set they remove some repeated data elements and removing un completed data elements. After that they selected the feature's from the data set . In this feature selection there are two main types of selection's are present, one is forward feature selection and second one is backward feature selection. After that the data set undergoes dimensionality reduction .After that they classify the data set into two parts one for training and one for testing . Than they train the both models such that are Random forest and support vector machine .Than they analysis the both models. After analysis they got 83% of accuracy for the support vector machine and 81.4 % for the support vector machine[21].

Proposed method

Data collection.

Diabetes dataset was obtained from the Pima Indian Diabetes dataset. It has 8 attributes namely Glucose,Blood Pressure,Skin Thickness,Insulin,BMI,Diabetes,Pedigree,Function,AgeOutcome
as shown in the Figure 1.
Identifying and saving the features that are available in the data set in features_list. Printing the columns name of the dataset as shown in the table 1.Setting the null value for each feature that is available in the data set.Printing the null values for each feature in the data set(This may use to get the duplicate elements in the data set and data preprocessing) as shown in the Table 1.

Table 1. Null values of each features

Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Counting the duplicate elements from the data set.Dropping the duplicate elements from the data setdf.drop_duplicates(['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']).Dataset after removing the duplicate elements.After removing the data set we have 768 rows of data available.Here we successfully remove the duplicates and completed the pre processing of dataset

Multicollinearity

Multicollinearity, which is the most problematic when performing LR, is checked.When multicollinearity exists, the explanatory power of the model decreases, and the model breaks when other variables are added.After checking multicollinearity, features with value of 10 or higher are removed.There are several methods to remove multicollinearity. The main methods are PCA

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

and VIF. I will use VIF.Note that you have to remove them one by one using 'loop'. Remove one and check the VIF value again.Implementing the multicollinearity on the data set.Method fuction to calculate mean and standard deviation for the data item and return the.These are the mean values and standard deviation for every feature in the data set ,In this we don't consider the value of mean and standard deviation because it represents the outcome value for the data set which we use in predicting shown in the Table 2.

Table 2. Mean and Standard deviation Values

Features	Mean	Standard Deviation
Glucose	120.89	31.97
BP	69.11	19.36
ST	20.54	15.95
Insulin	79.8	115.24
BMI	31.99	7.88
DPF	0.472	0.3313
Age	33.24	11.76

Here the table represents the mean and standard deviation of the all features that are available in the data set. This mean and standard deviation are helpful us to calculate multicollinearity between the all features. This is very important method in the feature extraction.

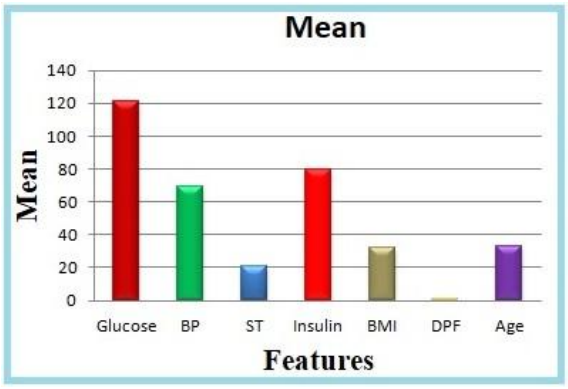

Figure 3.Features versus Mean

This graph represents the mean of every feature that we take as shown in the Figure 3. In this we can observe that glucose has highest mean value .next to that we have insulin and than Blood pressure. The DPF has the least mean because it either one or 0 only. So it mean value is between the 0 and 1.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

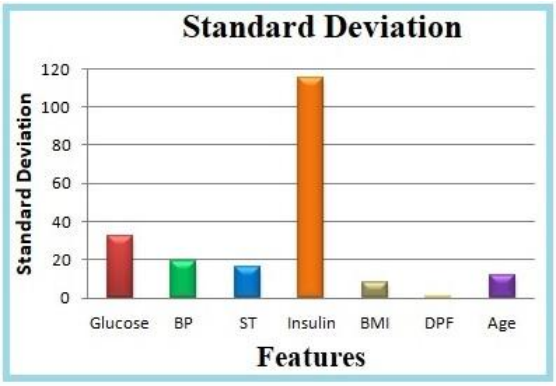

Figure 4. Feature versus Standard Deviation

This graph represents the mean of every feature that we take shown in the above Figure 4 . In this we can observe that insulin has the highest standard deviation .Than second glucose followe by blood pressure. The outcome has the lowest standard deviation ,because it contain only one or zero. .we finally return the sdf – Standard Deviation function value as a result for this function.That values of sdf are stored in the sdf.xlsx file. After adding every value to respected list its time to print the values. The output will be look like this.

Here we generate the sdf for every feature in the data set an check the correlation..Than we have to check the correlation value for every feature.Here we are check the correlation and store the values in the corr.xlsx file.. Creating a empty list's for the variables, regular coefficient and vif.VIF- variance inflation factor.Here variables are column names and vif is correlation at the specific point

TBable 4. Values of estimated using the mean and median values and VIF

Features	Estimate	VIF
Glucose	0.398	1.299
BP	-0.092	1.182
ST	0.0056	1.507
Insulin	-0.0506	1.424
BMI	0.22	1.297
DPF	0.0956	1.0641
Age	0.143	1.173

The table contains the values of estimated using the mean and median values and VIF as shown in the Table 4. Here the values of estimated may be negative also. The VIF value is useful us to do check the mutual collinearity between the different features.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

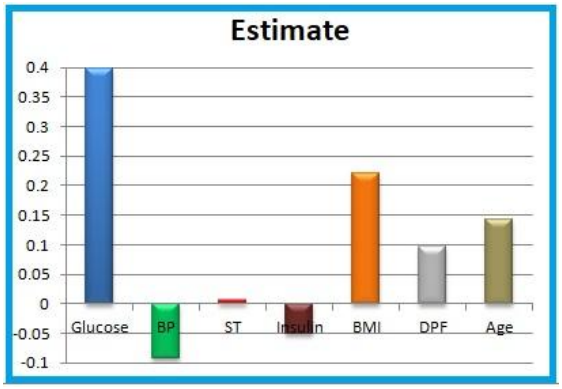

Figure 5. Values of estimated using the mean and median

Here the graph represents the estimated values of the all features in the data set as shown in the Figure 5 and Figure 6. The glucose has the highest estimated value and Blood pressure had the lowest estimated value , blood pressure and insulin has the negative estimated value and all remaining features has the positive estimated values.

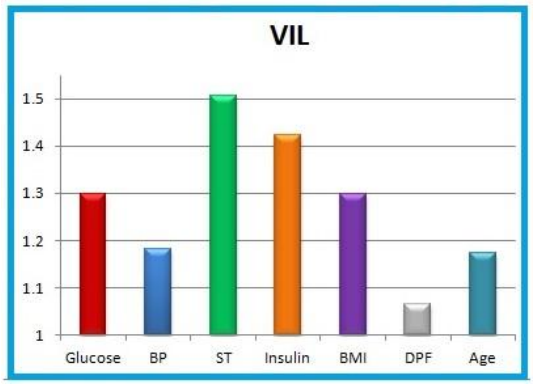

Figure 6. values of estimated using VIF

The VIF value is the final outcome of this test that represents the multi collinearity between the features. The feature that has ViF value more than 10 we have to eliminate that feature from the data set. We must select the select the remaining all the features that have VIF value less than 10.. But here in the given data set there is no feature whish value of the VIF value more than 10, So we cant eliminate any feature and we have consider all the features.

As we can see here there Is no value of VIF(variance inflation factor) is grater than the 10.So we can't eliminate any feature from the data set.So we have to use any other feature extraction method to extract the actual feature's from the data set.

HEAT MAP

Visualize and confirm the correlation between features. Although seemingly trivial, statistical analysis is a very important task.Heat map give you the correlation between the each feature on the other values that is same as correlation matrix in the multicolleaniarialy.This graph Shows just a visual representation of Correlation matrix shown in the Figure 7.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

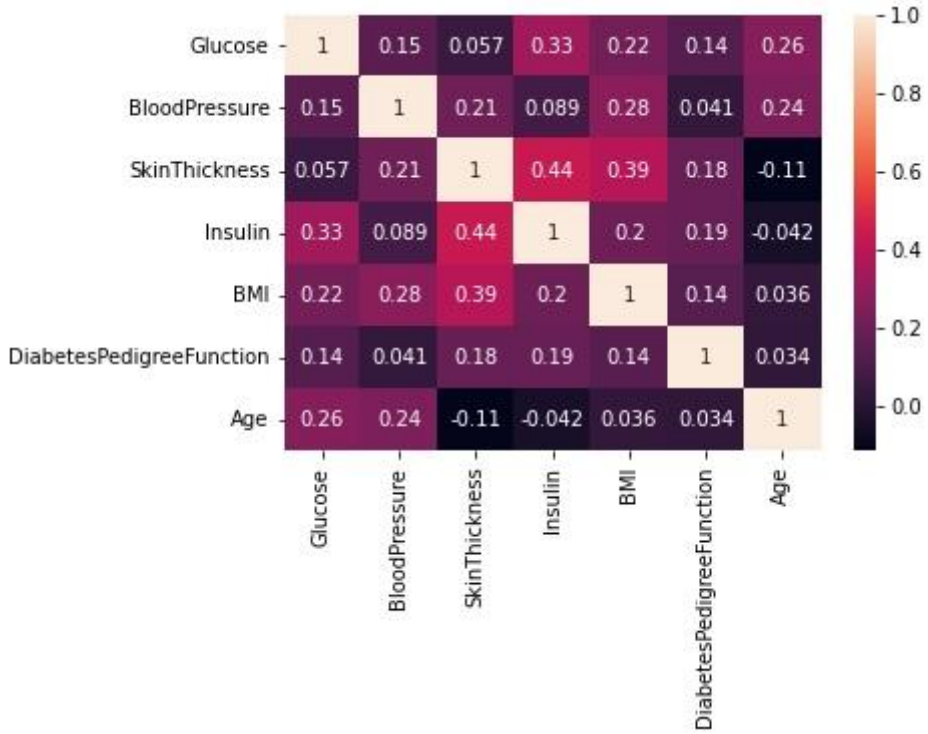

Figure 7. HeatMap

After taking the correlation values ,now it's time to go to T-Test

T-test

T-Test is nothing but a feature extraction method which is mainly depends on the value of p.We will first remove the variable by using the T-test.In the **T-Test** there are many factors that to be calculate that are :**F**= variation b/w sample means/variation within the sample.**Variable** – This is the variable for which the test was conducted.**Obs** – The number of valid (i.e., non-missing) observations used in calculating the t-test.**Mean** – This is the mean of the variable.**Std. Err.** – This is the estimated standard deviation of the sample mean.**Std. Dev.** – This is the standard deviation of the variable.

```
model2 = ols('Outcome ~ Glucose+BloodPressure+SkinThickness+Insulin+BMI+DiabetesPedigreeFunction+Age', df).fit()
table2 = sm.stats.anova_lm(model2, type=2)table2
```

Table 5.T-test

	df	sum_sq	mean_sq	F	PR(>F)
Glucose	1.0	37.983801	37.983801	232.525439	5.272565e-46
BloodPressure	1.0	0.006706	0.006706	0.041054	8.394876e-01
SkinThickness	1.0	0.442541	0.442541	2.709104	1.001905e-01
Insulin	1.0	0.494055	0.49405	3.024454	8.242281e-

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

			5		02
BMI	1.0	6.708843	6.708843	41.069528	2.576181e-10
DPF	1.0	1.651513	1.651513	10.110069	1.534422e-03
Age	1.0	3.043190	3.043190	18.629494	1.797405e-05
Residual	760.0	124.148517	0.163353	NaN	NaN

After Calculation of PR Value ,It's time to extract the features. Now we take the values only for which PR value is less than 0.005.df2 = table2[table2['PR(>F)'] < 0.05].df2 shown in the Table 6.After applying the T-Test we are come across few features. That are after extracting the features the data set is like below shown in the Table 6..features = df[['Glucose', 'SkinThickness', 'BMI', 'DiabetesPedigreeFunction']].Y = df['Outcome']

Table 6. Feature Extraction

Glucose	**SkinThickness**	**BMI**	**Diabetes**	**PedigreeFunction**
0	148	35	33.6	0.627
1	85	29	26.6	0.351
2	183	0	23.3	0.672
3	89	23	28.1	0.167
4	137	35	43.1	2.288
---	---	---	---	---
763	101	48	32.9	0.171
764	122	27	36.8	0.340
765	121	23	26.2	0.245
766	126	0	30.1	0.349
767	93	31	30.4	0.315

[768 rows x 4 columns]

After T-Test we come across through the only few features that are Glucose, SkinThickness, BMI, DiabetesPedigreeFunction..Now we have to prepare the data set to training and testing purposes train_features, test_features, train_labels, test_labels = train_test_split(features, Y).Here we using the train_test_split method we can get the all train_features, test_features, train_labels, test_labels using data set and features train_features

Preparing the data set for training testing the model.The labels are nothing but outcome values in the data set that is either 0 or 1. Here a scalar object can be created for the train features and test features.Now the data set is

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452
completely ready for comparing and testing using different algorithm.

Testing

First create an empty dictionary to save the accuracy values of the every algorithm that is tested using the extracted features

accuracy={}

accuracy is a dictionary which is used to store the accuracy values for each algorithm.

we prepared the data set to test the models ,so its time to test the each model to get the most accurate model from the testing of the different models.

Linear Regression

Here we used sklearn package which consists of LinearRegression.create an object name model which points to linear regression .Than fit the train_features, train_labels into the cerated object model to calculate the accuracy

The complete LinearRegression() is created now.For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to the created dictionary called accuracy.Print the accuracy with respect to algorithm name.LinearRegression: 0.2856586625004398

Logistic Regression

I wanted to use Backward Elimination, but gave up because the number of features was too small.

Here we used sklearn package which consists of LogisticRegression.create an object name model which points to LogisticRegression.Than fit the train_features, train_labels into the cerated object model to calculate the accuracy

The complete LogisticRegression()is created now.For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to the created dictionary called accuracy.Print the accuracy with respect to algorithm name.LogisticRegression: 0.75

K Nearest Neighbors

Here we used sklearn package which consists of ARDRegression.create an object name model which points to ARDRegression.Than fit the train_features, train_labels into the cerated object model to calculate the accuracy .The complete KNeighborsClassifier()is created now.For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to the created dictionary called accuracy.Print the accuracy with respect to algorithm name.K Nearest Neighbors: 0.8177083333333334

Decision Tree

Here we used sklearn package which consists of ARDRegression.create an object name model which points to ARDRegression.Than fit the train_features, train_labels into the cerated object model to calculate the accuracy

The complete DecisionTreeClassifier(criterion='entropy', max_depth=5)is created now.For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to the created dictionary called accuracy

Print the accuracy with respect to algorithm nameDecision Tree: 0.8177083333333334

Support Vector Machine

Here we used sklearn package which consists of ARDRegression.create an object name model which points to ARDRegression.Than fit the train_features, train_labels into the cerated object model to calculate the accuracy

The complete SVC()is created now .For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

the created dictionary called accuracy.Print the accuracy with respect to algorithm nameSupport Vector Machine: 0.796875

Naïve Bayes

Here we used sklearn package which consists of GaussianNB.create an object name model which points to GaussianNB.Than fit the train_features, train_labels into the cerated object model to calculate the accuracy

The complete GaussianNB()is created now.For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to the created dictionary called accuracy.Print the accuracy with respect to algorithm nameNaïve Bayes: 0.7708333333333334

Random Forest

Here we used sklearn package which consists of RandomForestRegressor.create an object name model which points to RandomForestRegressor.Than fit the train_features, train_labels into the cerated object model to calculate the accuracy.The complete RandomForestRegressor() is created now.For the model apply the score method to calculate the accuracy of the algorithm.add the accuracy to the created dictionary called accuracy.Print the accuracy with respect to algorithm nameRandom Forest: 0.895778558957048.After testing we have to take the one algorithm with highest accuracy score

Accuracy of ML Algorithms

The dictionary accuracy store all the related scores of all algorithm

Accuracy

Printing the accuracy dictionary shown in the Table 7 and Figure 8.

Table 7. Classification Accuracy

Methods	Classification Accuracy
Lin R	28.57
Log R	77.78
kNN	81.77
DTree	81.77
SVM	79.69
NB	77.08
RF	89.58

Here the table represents the accuracy of different models with the accuracy. The accuracy of linear regression is very low and below 50%. So this model wont help us to predict the diabetes. The remaining all models have accuracy levels more than 75% accuracy. The models k-nearest neighbour's, decision tree and Random forest have the accuracy more than the 80% .

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

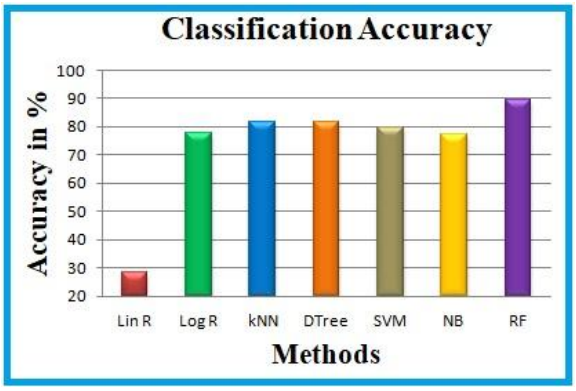

Figure 8. Classification Accuracy

The above graph represents the classification accuracy [22,23] of the different methods. In all methods the linear regression have the lowest accuracy where as all other features has nearly same accuracy. The Random Forest has the highest accuracy than any other methods. So we consider the random forest to train the model.plotting the bar graph for the each algorithm based on their accuracy values.Bar graph represents different algorithm scores.Now we have to find the maximum scored algorithm m=0algorithm=".Declered an empty algorithm and maximum score for comparing pourposes.This for loop take the all pairs of algorithm and value from the accuracy .Than it compere with the m value and store the highest score algorithm in the algorithm variable print("The algorithm which provides highest algorithm is",algorithm,"with accuracy ",m).printing the highest accuracy algorithm with the algorithm name and accuracy.The algorithm which provides highest algorithm is RF with accuracy 0.895778558957048.Now its time for implementing the algorithm which got highest score.Implementing the Random Forest algorithm

RANDOM FOREST IMPLEMENTATION

Creating the method random forest to implement and train that model. predict method in the sklearn will predict outcome the value for input values. After training the model using the predict method we can predict the output values..returning the outcome value that is predicted by training the model.

Input from user

Here we create the user_input list to store the input values that are given by the user. This while loop will help us to take exactly 10 numbers as input.. if user enter an invalid number than this loop will not take that input ,rather it will ask user to enter a correct 10 digits whats app number.

function calling in this section we will take the required features from the user entered features and assign it to the user_input

a=user_input

Here we call the function random forest to predict the output for specific input and we store it in the result variable

result =random_forest(user_input)

result

Sending message

Sending message is done by the pywhatkit module shown in the Figure 9 and Figure 10. Using this module we can send the messages through the sending message to receiver using pywhatkit except handling exception and printing error message

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 780-796
https://publishoa.com
ISSN: 1309-3452

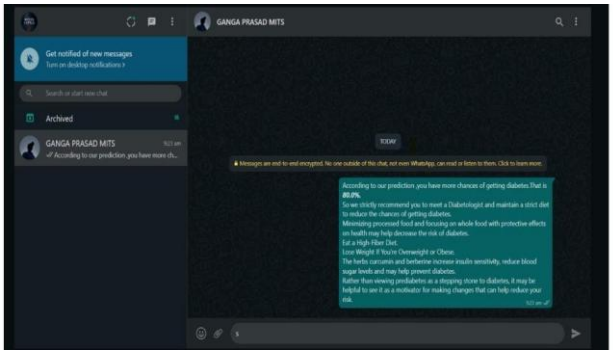

Figure 9. Message to the Laptop / Desktop / IoT devices

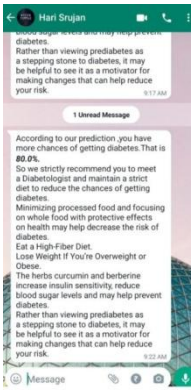

Figure 10. Message to the Whatsapp

Conclusion

From the data tabulated above , The Lin R approach show an classification accuracy (CA) of 28.57, The Log R approach show an classification accuracy (CA) of 77.78, The kNN approach show an classification accuracy (CA) of 81.77, The D Tree approach show an classification accuracy (CA) of 81.77, The SVM approach show an classification accuracy (CA) of 79.69, The NB approach show an classification accuracy (CA) of 77.08, The RF approach show an classification accuracy (CA) of 89.58. From this, we can infer that for diabetic the RF approach gives the best performance compared to all other approaches. In future we can enhance the research using medical cell image diabetics databases and also with the recent technology like IoT, cloud computing, Block chain technology we can develop a smart portable automatic early detection of diabetics device with secured storage.

References

1. P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, “Machine Learning and the Prediction of Hydrocephalus,” JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.
2. “The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company.” [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-andservices/our-insights/the-big-data-revolution-in-us-health-care. [Accessed: 12-May-2018].
3. A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, “Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing,” IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.

4. L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350–361, May 2017.
5. J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.
6. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.
7. K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings," IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016.
8. K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization Based on Social Big Data Analysis in the Vehicular Networks," IEEE Trans. Ind. Informatics, vol. 13, no. 4, pp. 1932– 1940, Aug. 2017.
9. E. Ahmed et al., "The role of big data analytics in Internet of Things," Computer Networks, vol. 129, no. December, pp. 459–471, 2017
10. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.
11. A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big Data Analytics in Healthcare," Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.
12. J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big Data for Health," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015.
13. Alshamlan, Hala; Taleb, Hind Bin; Al Sahow, Areej (2020). [IEEE 2020 11th International Conference on Information and Communication Systems (ICICS) - Irbid, Jordan (2020.4.7-2020.4.9)] 2020 11th International Conference on Information and Communication Systems (ICICS) - A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression. , (), 1–4. doi:10.1109/ICICS49469.2020.239549.
14. Usman, Sahnius; Reaz, Mamun Bin Ibne; Ali, Mohd Alauddin Mohd (2016). [IEEE 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) - Malaysia (2016.12.4-2016.12.8)] 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) - Risk prediction of having increased arterial stiffness among diabetic patients using logistic regression. , (), 699–701. doi:10.1109/IECBES.2016.7843540
15. Chaudhuri, A. K., & Das, A. (2020). Variable Selection in Genetic Algorithm Model with Logistic Regression for Prediction of Progression to Diseases. 2020 IEEE International Conference for Innovation in Technology (INOCON). doi:10.1109/inocon50539.2020.9298372.
16. Ahmed, Hager; Younis, Eman M.G.; Ali, Abdelmgeid A. (2020). [IEEE 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE) - Aswan, Egypt (2020.2.8-2020.2.9)] 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE) - Predicting Diabetes using Distributed Machine Learning based on Apache Spark*. , (), 44–49. doi:10.1109/ITCE48509.2020.9047795.

17. Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2021). A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic. IEEE Access, 9, 7869–7884. doi:10.1109/access.2020.3047942
18. Barhate, Rahul; Kulkarni, Pradnya (2018). [IEEE 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) - Pune, India (2018.8.16-2018.8.18)] 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) - Analysis of Classifiers for Prediction of Type II Diabetes Mellitus. , (), 1–6. doi:10.1109/ICCUBEA.2018.8697856
19. P, Anirudh Hebbar; M V, Manoj Kumar; H A, Sanjay (2019). [IEEE 2019 1st International Conference on Advances in Information Technology (ICAIT) - Chikmagalur, India (2019.7.25-2019.7.27)] 2019 1st International Conference on Advances in Information Technology (ICAIT) - DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes. , (), 271–276. doi:10.1109/icait47043.2019.8987277
20. VijiyaKumar, K.; Lavanya, B.; Nirmala, I.; Caroline, S. Sofia (2019). [IEEE 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) - Pondicherry, India (2019.3.29-2019.3.30)] 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) - Random Forest Algorithm for the Prediction of Diabetes. , (), 1–5. doi:10.1109/ICSCAN.2019.8878802.
21. Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). doi:10.1109/icaccs51430.2021.9441935.
22. Bose S.C., Veerasamy M., Mubarakali A., Marina N., Hadzieva E. (2020) Analysis of Feature Extraction Algorithm Using Two Dimensional Discrete Wavelet Transforms in Mammograms to Detect Microcalcifications. In: Smys S., Tavares J., Balas V., Iliyasu A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer, Cham. https://doi.org/10.1007/978-3-030-37218-7_4.
23. Bose, J.S.C., Shankar Kumar, K.R.: Detection of micro classification in mammograms using soft computing techniques. Eur. J. Sci. Res. 86(1), 103–122 (2012)