A Comparative Accuracy Prediction Analysis of Diabetes Disease by Using Machine Learning Algorithms

Nikhil Sharma[1], Savita[2]
[1]Research Scholar, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India

[2]Assistant Professor, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India

Abstract: Diabetes is a critical disease that many people suffer from worldwide. It affects the human body's fuel systems. People who are suffering from diabetes have a very high risk of developing diseases like heart disease, kidney disease, stroke, eye problems, nerve damage, etc Generally, eating food is converted into glucose and supplied to different parts of the body through blood. The insulin release by the pancreas helps to balance blood glucose levels. Due to an increased level of glucose in the body, our body stops the production of insulin, which is the leading cause of diabetes. According to the 2022 survey report of t h e Diabetes Federation, 500 million people are presently living with diabetes worldwide. By 2040, this will reach up to 750 million, 01 out of 10 will be living with diabetes. Disease. The research's main objective is to develop an intelligent system that helps in the early prediction of Diabetes and also helps in making correct, accurate, and timely predictive decisions. The key value features are selected by using Pearson's correlation coefficient method. Various machine learning classifiers are used to predict the disease. The random forest algorithm achieved the highest accuracy of 0.83, and Linear regression achieved the lowest accuracy score of 0.64.

Keywords: Classification models, supervised machine learning, Random Forest, Linear Regression, SVM

INTRODUCTION

Diabetes is a metabolic disorder that generally affects glucose production, and glucose is the primary energy source for the body's cells. Our body consumes glucose from food and transfers it to various body cells With the help of the bloodstream. The pancreas creates insulin, which helps to allow glucose to enter cells. A body of diabetes-affected patients stops producing insulin, so glucose starts accumulating in the bloodstream and causes several dangerous problems like heart problems, kidney problems, blindness problems, etc. It can be easily managed easily by taking medicine and healthy foods. Regular body exercise also helps to balance blood sugar levels. Generally, diabetes has two types. Type 1 diabetes typically affects the body's immune system, so our body stops insulin production in the body. Children and younger generations are mostly affected by Type 1 diabetes [1][2]. Type 2 is another type of diabetes, which is caused by a lack of insulin production by the pancreas, making it difficult for glucose to enter the cells. Obesity, less body exercise, and high consumption of a fatty diet are the most common causes of type 2 diabetes, and overweight people are more likely to develop type 2 diabetes. Sometimes family history also, in some cases, plays a significant role in the causes of diabetes. If someone in your family, a parent or sibling, has diabetes, the probability of developing diabetes is increased [3]. Diabetes symptoms initially develop slowly with time, and sometimes patients are not aware that they are affected by diabetes. Some common indications are increased thirst levels, and our body needs more liquid. In diabetes patients, the frequency of urination also increases, and the frequency is higher during night time. The second most common symptom is more hunger The body demands more food for the consumption of energy [3]. Fatigue is the third most important symptom of diabetes, in which diabetes people with diabetes may feel more weakness and tiredness, effects on eye vision, and take more time to heal cuts in their body [4][5].

PROBLEM DEFINITION AND METHOD

The research paper's main objective is to develop an intelligent system that provides a timely and accurate prediction of diabetes disease. The research paper concluded the comparative accuracy prediction analysis of diabetes disease by using different machine learning classifier models of supervised machine learning. [6][7][8]. The designed model helps select the best models for predicting diabetes. Apart from that, it also helps to identify problems and their recovery solutions promptly and increases patients' life expectancy. Problem definition identifies the impact and importance of research and provides timely healthcare solutions to problems. Research work includes data collection, feature selection, data pre-processing, data training and testing, model selection, accuracy, and future scopes.

A. Data Collection Preparation

Data collection is the primary initial step for designing any model. Correct and authentic data are crucial in achieving the best accuracy from designed models. Data is extracted from the Kaggle online repository [9][10]. The dataset has 9 columns and 768 rows. After feature selection, only seven columns are selected as input features to train the model. A detailed description of the dataset with all its features is given in Fig. 1

Fig. 1. Feature Description

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0

B. Data Preprocessing

After data collection, data preprocessing is the next most crucial step in designing a model. Data needs to be filtered and cleaned, and various data cleaning and scaling techniques are applied to clean and scale the data. The data cleaning process depends on the size of the dataset. If the dataset size is large, remove missing rows from the dataset; if the dataset size is small, then replace missing values either by the average value or the mean value.

```
a.isna().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Fig. 2. Missing value Information

The second data cleaning process is scaling. Scaling is used when there is a significant variation between the values of columns. It scales values between zero and one. After the data cleaning, data is divided into train and test parts. Dataset does not have any missing values, as shown in Fig. 2.

C. Feature Selections

Features are the main backbone of any model, and accurate feature selection is the most crucial step for the designer; selecting the wrong features increases the chances of overfitting and under fitting. A wrong selection of features may affect the accuracy of models. In machine learning, the heat map technique is mainly used to select the correct features. A heat map generally selects the features by checking the dependency between the input variables, and it finalizes the key features for model training, which are highly dependent on each other. The value of the heat map always lies between 0 and 1. If the value is one, it means features are highly reliant on each other, and if the value is 0, features are less dependent on each other[11][12]. The heat map is shown in Figure 3

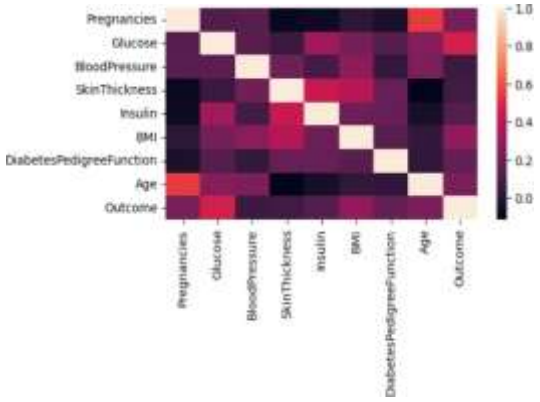

Fig. 3. Heat Map

MODEL TRAINING & SELECTION

In Model Training, the dataset is divided into two sub-parts- training and testing. The first large part is supplied for the training model and the remaining. small part is used for testing purposes. In the proposed model, the size of the dataset is not too large, so 90% of the data is fed to the model for training, and the remaining 10% of the data is for testing purposes. Sometimes, it has been analyzed that the accuracy of the designed model is very high when testing trained data values and very low when testing data. This problem arises due to the model being over fitted. To find the best accuracy of the model, it should be highly accurate for testing data. After training and testing, the next most important step is the selection of a classifier, and the proper choice of a correct model has a high impact on the accuracy of model. next most important step is the selection of a classifier, and the proper choice of a correct model has a high impact on the accuracy of model [13][14]. In the proposed methodology, various supervised machine- learning classifiers are used to classify or for the prediction of Diabetes. The overview of classification algorithms as follows: small part is used for testing purposes.]. In the proposed methodology, various supervised machine- learning classifiers are used to classify or for the prediction of Diabetes. The overview of classification algorithms as follows:

Linear Regression

Linear regression is a straightforward and most demanded algorithm of machine learning. Generally, it checks the relationships among independent variables and dependent variables. It is generally used to predict continuous numeric values. It is mostly used to predict sales, salary, age, product price, etc. In this method, the value of the dependent variable varies with a change in the values of the independent variables. [5][9][10]. The Implementation of the Model is shown in Figure 4.

```
#Train Test Model
from sklearn.model_selection import train_test_split
b_train, b_test, c_train, c_test = train_test_split(b, c, test_size=0.1)
model=LinearRegression()
model.fit(b_train,c_train)
model.predict([[77,92,1.015,3.0,2,105,35.0,2.5,145.0,4.2,9.1,33,7800,3]])
```

Fig 4. Implementation of Linear Regression

Support Vector Machine (SVM)

SVM is the most famous, demanded, and popular supervised learning algorithm. It can be used for both types of problems: classification and regression. But it provides a good, accurate result for the prediction of classification problems. The SVM model creates a decision boundary line so that new data can be easily classified into the right classes. A hyperplane represents the decision boundary. Vector points that are used to create decision boundary lines and these points are known as margin points or support vectors [7][13[15]. The implementation work is shown in fig 5.

```
# Support Vector Machine Using Train Test Model
from sklearn.model_selection import train_test_split
b_train, b_test, c_train, c_test = train_test_split(b, c, test_size=0.1)
model2=SVC(gamma=5)
model2.fit(b,c)
model2.predict([[77,92,1.015,3.0,2,105,35.0,2.5,145.0,4.2,9.1,33,7800,3]])
```

Fig 5. Implementation of SVM Model

Random Forest

Random Forest is generally used for large data sets. It contains multiple decision trees, and the final predictive result is calculated by taking the average values to predict the output by using majority voting. It is mostly used for classification problems but can also be used for regression problems. It is based on the concept of ensemble learning and solves complex problems by using different classifiers to improve the model's performance. The implementation work is shown in fig 6.

```
#train test random forest model
from sklearn.model_selection import train_test_split
b_train, b_test, c_train, c_test = train_test_split(b,c,test_size=0.1)
model2=RandomForestClassifier(n_estimators=10,n_jobs=-1)
model2.fit(b_train,c_train)
model2.predict([[77,92,1.015,3.0,2,105,35.0,2.5,145.0,4.2,9.1,33,7800,3]])
```

Fig 6. Implementation of Random Forest

Decision Tree

Decision Tree is also used to solve both classification and regression problems. A decision tree behaves like a tree where the root node is divided into several branches. The tree is developed in which the internal nodes represent the features of the data set, and the leaf nodes represent outcomes. It uses the CART algorithm to build a tree, which is further split into subtrees based on the answer Yes/No. The implementation work is shown in fig7

.

```
#Decision Tree with Train test model
from sklearn.model_selection import train_test_split
b_train, b_test, c_train, c_test = train_test_split(b, c, test_size=0.1)
dtree1=tree.DecisionTreeClassifier()
dtree1.fit(b_train,c_train)
dtree1.predict([[77,92,1.015,3.0,2,105,35.0,2.5,145.0,4.2,9.1,33,7800,3]])
```

Fig 7. Implementation of Decision Tree

ACCURACY ANALYSIS

Various scored metrics, such as accuracy, precision, recall, and F1-score measure the accuracy of all machine learning classifiers. The proposed model also used the K-Fold cross-validation approach for calculating the best accuracy of the model. Prediction analysis of various classification algorithms is given in Table 1 to Table 4

TABLE 1: Prediction Analysis of SVM

Classification Report			
	precision	recall	f1-score
0	0.72	0.70	0.71
1	0.74	0.76	0.75
weighted avg accuracy	0.73	0.73	0.73

TABLE 2: Prediction Analysis of Decision Tree

Classification Report			
	precision	recall	f1-score
0	0.74	0.76	0.75
1	0.82	0.80	0.81
weighted avg accuracy	0.78	0.78	0.78

TABLE 3 : Prediction Analysis of Linear Regression

Classification Report			
	precision	recall	f1-score
0	0.63	0.61	0.62
1	0.65	0.67	0.66
weighted avg accuracy	0.64	0.64	0.64

TABLE 4: Prediction Analysis of Random Forest

Classification Report			
	precision	recall	f1-score
0	0.81	0.82	0.82
1	0.85	0.84	0.84
weighted avg accuracy	0.83	0.83	0.83

From comparative accuracy performance analysis of various machine learning classifiers, it has been observed that the random forest algorithm scored the maximum accuracy score of 0.83 among all classification algorithms. Linear regression has achieved the minimum accuracy score, with a value of 0.64. Table 5 represents the accuracy analysis of applied classification algorithms.

Table 6. Accuracy Comparision

S.No	Models	Accuracy
1	SVM	0.73
3	Random Forest	0.83
4	Decision Tree	0.78
5	Linear Regression	0.63

Graphical representation of the accuracy scores of classification models is given in Figure 6.

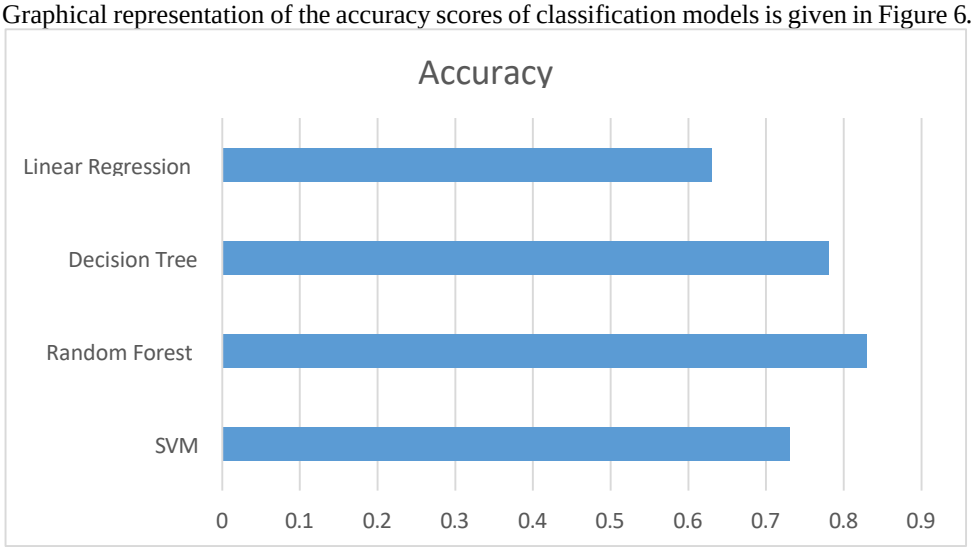

Fig. 6 . Accuracy Scores

RESULT AND CONCLUSION

The proposed model provides the most accurate model for the prediction of diabetes. Diabetes is a common disease, and early treatment is beneficial for reducing its effects. Early treatment also helps in finding the appropriate causes or reasons for the occurrence of the disease, and timely and suitable treatment increases life expectancy. From related research, it has been observed that design models provide an accurate or correct predictive model for early prediction of disease and help in making the right appropriate decision regarding understanding the cause of diabetes. Research also classified the strengths and weaknesses of various designed models and helped to identify the best model based on the accuracy score and efficiency of the model. Ma. After the comparative study of various designed models, it has been found that the random forest model has high efficiency or accuracy among all models, and the linear regression model has the minimum accuracy. In the future, large datasets can be used for the prediction of diabetes disease, and some deep learning algorithms can also be used for early prediction.

REFERENCES

[1].Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*. 1997;20:1183– 97.

[2] Norris SL, Lau J, Smith SJ, Schmid CH, Engelgau MM. Self- management education for adults with type 2 diabetes:A meta-analysis of the effect on glycemic control. *Diabetes Care.* 2002;25:1159–71.

[3] Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Pract.* 2010;87:4–14.

[4] Pradeepa R, Deepa M, Datta M, Sudha V, Anjana RM Unnikrishnan R, et al. Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India:Phase I results of the Indian Council of Medical Research India Diabetes (ICMRINDIAB) study. *Diabetologia.* 2011;54:3022–7.

[5] Ramachandran A, Snehalatha C, Salini J, Vijay V. Use of glimepiride and insulin sensitizers in the treatment of type 2 diabetes-a study in Indians. *J Assoc Physicians India.* 2004;52:459–63.

[6] Wagai GA, Romshoo GJ. Adiposity contributes to poor glycemic control in people with diabetes mellitus, a randomized case study, in South Kashmir, India. *J Family Med Prim Care.* 2020.

[7] . Birjais R, Mourya AK, Chauhan R, Kaur H. Prediction and diagnosis of future diabetes risk:A machine learning approach. *SN Appl Sci.* 2019;1:1–8.

[8] Sadhu A, Jadli A. Early-stage diabetes risk prediction:A comparative analysis of classification algorithms. *Int Adv Res J Sci Eng Technol (IARJSET)* 2021;8:193–201

[9] Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. *J Phys Conf Ser*. 2020;1684:1–6.

[10] TM, Vo TM, Pham TN, Dao SV. A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. *IEEE Access.* 2020;9:7869–84.

[11] Julius AO, Ayokunle AO, Ibrahim FO. *Early diabetic risk prediction using machine learning classification techniques.* Available from.

[12] Shafi S, Ansari GA. Early prediction of diabetes disease &classification of algorithms using machine learning approach. *In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)* Available from:SSRN 3852590

[13] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *ICT Express.* 2021;7:432– 9

[14] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia Comput Sci.* 2018;132:1578– 85.

[15] Agrawal P, Dewangan AK. A brief survey on the techniques used for the diagnosis of diabetes-mellitus. *Int Res J Eng Tech IRJET.* 2015;2:1039–43.