Diabetes Prediction Using Supervised Machine Learning Models

V.Manohar Chand Naik
Department of Electrical and Electronics Engineering
NIT Nagaland
Chumukedima – 797 103, India
manoharchandnaik@gmail.com

M.Prakash
Department of Electrical and Electronics Engineering
NIT Nagaland
Chumukedima – 797 103, India
prakash@nitnagaland.ac.in

B.Shakila
Department of Electrical and Electronics Engineering
NIT Nagaland
Chumukedima – 797 103, India
shakila@nitnagaland.ac.in

2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON) | 979-8-3503-3355-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/INDISCON58499.2023.10270102

***Abstract*—Diabetes is regarded as one of the most chronic illness that raises blood sugar level. Versatile disorders including kidney failure, variation in blood pressure, eye defectiveness, heart problems and other major organ failures are directly related to undiagnosed/untreated diabetes. Early identification of diabetes assisted by machine learning techniques shall provide definite health tracking even before the organ failures. The present study focuses on creation of a machine-learning model that can predict a patient's chances of developing diabetes, with the highest degree of accuracy. Two machine learning classification methods: K-Nearest Neighbors (KNN) and Logistic Regression (LR) have been proposed in this study for Pima Indians Diabetes Dataset (PIDD), to identify the diabetes at an early stage with greater accuracy. The effectiveness of the proposed algorithms were assessed using various metrices including confusion matrix, recall, precision, accuracy, and F_1-measure. Accuracy is evaluated in a precise way after obtaining two distinct dataset: valid and erroneous. Results obtained through LR surpasses other algorithms with the maximum accuracy of 84.89%.**

***Keywords*— Machine Learning Predictions, Supervised Learning models, Logistic Regression, K-Nearest Neighbors, Classification problems and Diabetes Prediction.**

I. INTRODUCTION

One of the chronic (long-lasting or constantly recurring) diseases in the world is diabetes. Diabetes Mellitus (DM), a category of metabolic illnesses is characterized by impaired insulin secretion, Hyperglycemia (high blood sugar), poor protein and carbohydrate metabolism caused by insulin insufficiency. In general, three types of diabetes: TYPE 1, TYPE 2 and GESTATIONAL diabetes are widely considered. An abnormal high blood sugar level is caused by type 1 diabetes, due to the body's inability to manufacture the hormone "Insulin". The insulin secretion cells in the pancreas are further attacked by the immune system of the body. A necessary duty is completed by insulin that it makes it possible for the blood glucose to enter into cells and nourish human body. Glucose level induction from the carbohydrates available in food and drink will continue to prevail, even if a person is identified with type 1 diabetes. When glucose enters into bloodstream, it cannot enter the body's cells without insulin. This causes a build-up of glucose in blood, which in turn escalates blood sugar level.

In type 2 diabetes, the pancreas fails to produce an appropriate proportion of insulin. As a result, one's blood sugar level keeps rising. 90% of people from United Kingdom (UK) have been identified with type 2 diabetes [6]. Several organs including feet, heart and eyes can be seriously affected by excessive blood sugar levels, especially in case of type 2 diabetes. These are referred to as complications of diabetes. However, with the right care and medication, there are chances to reduce the probability of getting type 2 diabetes.

Throughout pregnancy, gestational diabetes can become apparent. It affects women who never had diabetes before. Further it is suggested that, one who have high blood sugar level needs to take extra care of themselves as well as the developing child. Healthy food and exercise are mandates to achieve the same. After delivery, sugar level usually gets back to normal scale. Blood test is done between 24 and 28 weeks of pregnancy to identify the gestational diabetes at early stage.

Several research works were carried out to understand the performance of various machine learning algorithms in predicting and analyzing diabetes, with implications for early detection and prevention of the risk caused by diabetes. Four classification algorithms including Decision Tree, ANN, Naive Bayes and SVM were considered for early detection of diabetes [1]. Precision rate of 77.3% was achieved with SVM based classification algorithm. Applications of six machine learning algorithms for predicting diabetes using a dataset of patients' medical records were discussed [2]. SVM and KNN algorithms provide the highest accuracy of 77% for predicting diabetes where LR gives 74%. The study concludes that predictive analytics in healthcare can improve medical decision-making, but limitations in the dataset size and missing attribute values impact accuracy. Another study [3] evaluated three algorithms' performance in which the Naive Bayes outperformed the other two algorithms with an accuracy of 76.30%. Random Forest was found to be the best algorithm for predicting diabetes in a review of crucial factors contributing to diabetes [4]. Literature review [5] focuses on the effectiveness of predicting diabetes among Indian pregnant women. The study emphasizes the importance of processing medical data for analysis purposes, with potential applications in various medical sectors. Further researchers can explore the feasibility of using feature selection and deep neural networks to improve the accuracy of the model.

979-8-3503-3355-8/23/$31.00 ©2023 IEEE

Diabetes is a serious disease causing multiple complications, with an increasing global prevalence expected to reach 700 million by 2045 [12]. Early detection is crucial. K-Nearest Neighbors and Naive Bayes algorithms were adopted to predict diabetes using the Pima Indians dataset. Naive Bayes outperforms KNN with an average accuracy of 76.07%, precision of 73.37%, and recall of 71.37%, whereas KNN has an average accuracy of 73.33%. The study suggests using Naive Bayes algorithm for predicting diabetes.

II. MACHINE LEARNING AND ITS TYPES

Machine Learning (ML), an integral part of Artificial Intelligence (AI) is presumed to enable the computers to "Learn" by themselves from training dataset (by feeding) and enhances its performance over time without needing to be explicitly programmed by humans. Machine learning algorithms shall develop the predications through pattern recognition and consistent learning. Machine learning can make use of enormous volumes of data and is far more accurate than humans and thus saves money and time. Machine learning is classified into three types as follows.

A. Supervised learning

Predictions are made by supervised learning algorithms and models, using labelled training data. Both the desired result and set of inputs are included in each training sample. The sample data examined by a supervised learning algorithm draws an inference, a random guess about the labels of unobserved data. The supervised learning strategy is the most typical and well-liked method of machine learning.

1. *Classification:*

Classification predicts either class or discrete values. Support Vector Classifier (SVC), Random Forest (RF), Decision Tree Classification, XGBoost classification, AdaBoost classifier etc., are few examples of classification algorithms.

2. *Regression:*

It is based on predicting the continuous value or quantity, where decimal values are involved in this method. Examples include prediction of salary, price, distance, temperature etc., Logistic Regression (LR), Decision Tree (DT) Regression etc., are classified under regression algorithms.

A. Unsupervised Learning

Unlabeled data are generally handled using unsupervised learning algorithms, in order to uncover patterns and correlations. Models are given input data and unknown output are left to conclude the pattern on their own without any training. Clustering and Association are the major classifications of unsupervised learning strategy.

1. *Clustering:*

Clustering assembles related data into groups and is one of the most used method categorized under unsupervised learning. Hidden trends or patterns can be identified using this strategy and is frequently employed for exploratory analysis. “k-means clustering” and “Hierarchical clustering” are renowned clustering algorithms.

2. *Association*:

Significant relationships between data points in datasets shall be discovered using unsupervised association methodology. For instance, if customer 1 (C_1) selects a list of items including A, B, C and D followed by customer 2 (C_2) with preference including A, B, E and F it is obvious that customer 3 (C_3) may prefer item B when he selects item A in his cart. It is decided based on the selection of C1 and C2. Primary application of this technique is evident as suggestion notifications on e-commerce sites.

B. Reinforcement Learning

Reinforcement learning (RL) focuses on maximization of rewards. In essence, reinforced machine learning algorithms look for the optimum course of action to follow in a certain circumstance. They experiment and learn as they go along. The model learns through its own errors and selects the behavior automatically. The learning and selection process yields either the best solution or the greatest reward, since there is no training data. Robotics and video games are two such major applications for this machine learning technique. Video games show a clear correlation between actions and outcomes and use scorekeeping to determine success.

C. Proposed Model for Diabetes Prediction

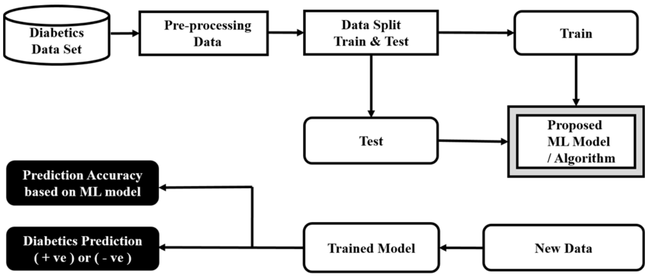

Diabetics Data Set
Pre-processing Data
Data Split Train & Test
Train
Test
Proposed ML Model / Algorithm
Prediction Accuracy based on ML model
Diabetics Prediction (+ ve) or (- ve)
Trained Model
New Data

Fig: 1 proposed model for predicting diabetes using ML

The PIDD was collected and pre-processed to eliminate missing values, noise and other inconsistencies before implementing the classification algorithms. After data preprocessing, the data set is separated and categorized as training dataset and testing dataset. The trained model will further classify new random data under designated label and the accuracy shall be observed alongside.

III. SUPERVISED MACHINE LEARNING MODELS

Model selection is known as the process of, selecting a model as the ultimate solution to the ML predictions. As the optimum model to solve this problem is unknown, it cannot be predicted which model will perform the best. Therefore, a variety of models that are fit for the predictions of diabetes are analyzed and assessed. The patterns buried in data are represented mathematically by a machine learning model. A controlling structure is found by the machine learning model when it is developed, fitted or trained using the training data. Rules have been written to represent this regulatory framework and can be used to anticipate outcomes in novel scenarios. In essence, the model will be able to identify some relation inside the new data after being trained on training data.

A. *Data Preprocessing*

Unbalanced and duplicate data, as well as features with a high correlation and minimal variation, and outliers are handled, and the data is standardized in the data pre-processing step. The model was developed using the analyzed data. Appropriate pre-processing and structuring of the data is necessary before applying classifiers. Special attention should be given to the data before fitting it to the model. To ensure accurate and reliable findings, inconsistent data is processed and removed as well as missing values should be detected and appropriately handled to provide more precise results. Following this, the input parameters and output column should be separated, and the data should be standardized for improved performance before being fed into the machine learning model (MLM).

B. *LOGISTIC REGRESSION MODEL*

LR model is used for both classification and regression in supervised learning. It is particularly effective for binary classification problems and is easy to implement. Additionally, it performs well on larger datasets that exhibit linear relationships. LR operates based on a sigmoid function. The sigmoid function is given as follows.

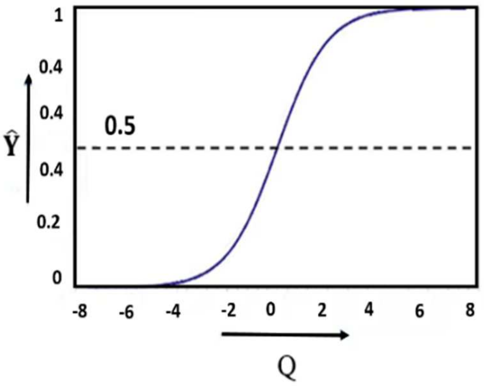


Fig: 2 Sigmoid Fuction Curve

$$\text{Sigmoid Function } (\hat{Y}) = \frac{1}{1+\exp^{-Q}} \qquad (1)$$

Where $\hat{Y}$ is probability percentage of the X data point for getting 1, if $\hat{Y}$ is more than 50%, outcome will be 1 and if $\hat{Y}$ is less than 50%, predicted outcome will be 0.

Where , ($\hat{Y}$) = Probability that (Y =1 | X)
exp = Exponential (e = 2.718281)
Q = is known as Line equation $(W_e).X + B_S$
(Where number of weights = Number of input features)
W_e = Weights of the inputs features
B_S = Bias
X = Input features of datasets

1. *Weights (W_e):*

There were three different kinds of weights including positive, negative and zero weights, which are also known as the slope of the line equation.

a) *Positive and Negative weight:* If the input feature has any direct proportion impact on the resultant output and positively correlated then that particular input feature contains positive weights. And vice versa for negative weight.

Ex: Sugar level and diabetes outcome (directly proportional).

b) *Zero weight:* The specific input characteristic with zero weight can be eliminated from the dataset if it is unreliable on resultant output.

Ex. Salary and the diabetes outcome (non-related input features).

2. *Bias (B_S):*

Bias is a constant value, also known as the intercept in line equation, is assigned a random value initially. When the bias is strong the model is said to be highly biased. However, as the bias decreases, the model's performance and accuracy improves, it can be achieved by turning the parameter.

3. *Types of Logistic Regression :*

a) *Binomial Logistic Regression:* In this LR, two potential results are available for the categorical response, such as Diabetes positive or negative, which are represented as 1 or 0.

b) *Multinomial Logistic Regression:* Three or more categories of outcomes are present here without any sequence or order. Predicting which meal will be more popular, for instance (Veg, Non-Veg, Vegan).

c) *Ordinal Logistic Regression:* A minimum of three ordered categories are present which are in order. Example: Predicting the House price based on area and crime rate.

4. *Decision Boundary*:

A limit can be set to determine the class to which a piece of data belongs. The estimated probability is categorized into groups based on the threshold. The decision boundary can be linear or non-linear, and the polynomial order can be increased to obtain more complex decision boundaries. For example, email can be classified as spam if the predicted value is less than 0.5.

5. *Loss Function (L):*

The mean squared error can be used as a cost function for linear regression, and in case of logistic regression, it will be a convex function of parameters. The global minimum can be achieved through gradient descent, and local minimum can be reached only if the function is non-convex. The equation is given as

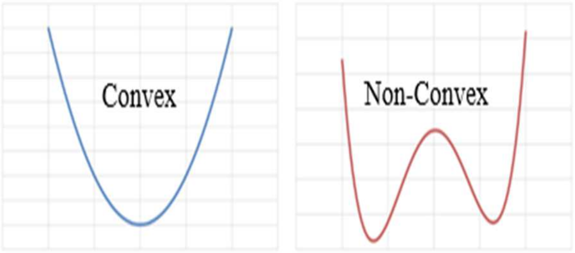


Fig: 3 Convex and Non-Convex Graph Representation

$$\text{LOSS} = \frac{1}{\text{n}} \sum_{j(it)=1}^{n} (\text{Y}_{j(it)} - \hat{\text{Y}}_{j(it)})^2 \quad (2)$$

$$\text{L} (\text{Y}_{(it)},\hat{\text{Y}}_{(it)}) = - (\text{Y}_{(it)} \log\hat{\text{Y}}_{(it)} + (1- \text{Y}_{(it)}) \log (1- \hat{\text{Y}}_{(it)})) \quad (3)$$

When $\text{Y}_{(it)}=1$;

$$\text{Then L} (1, \hat{\text{Y}}_{(it)}) = - (1 \log\hat{\text{Y}}_{(it)} + (1 - 1) \log (1 - \hat{\text{Y}}_{(it)}))$$

$$\text{L} (1,\hat{\text{Y}}_{(it)}) = - (\log\hat{\text{Y}}_{(it)}) \quad (4)$$

A smaller Loss function value is always desired, therefore $\hat{\text{Y}}_{(it)}$ should be made very large, so that (-log$\hat{\text{Y}}_{(it)}$) becomes a large negative number.

When $\text{Y}_{(it)}= 0$;

$$\text{Then L} (0,\hat{\text{Y}}_{(it)}) = - (0 \log\hat{\text{Y}}_{(it)} + (1 - 0) \log (1 - \hat{\text{Y}}_{(it)}))$$

$$\text{L} (0,\hat{\text{Y}}_{(it)}) = - \log (1 - \hat{\text{Y}}_{(it)}) \quad (5)$$

A smaller Loss function value is desired , hence $\hat{\text{Y}}_{(it)}$ should be very small, so that (- log (1 - $\hat{\text{Y}}_{(it)}$)) will be a large negative number .

6. Cost function (J):

Logarithmic loss, also known as log loss, is used in LR. Although raw log-loss statistics can be difficult to interpret, log-loss remains a valuable tool for comparing models. A lower Log loss value corresponds to better forecasts for any given situation. The negative mean of the log of adjusted projected probability for each case is referred as log loss. It deals with the batch as a whole or a punishment for a number of training sets.

$$\text{L} (\text{Y}_{(it)}, \hat{\text{Y}}_{(it)}) = - (\text{Y}_{(it)}\log\hat{\text{Y}}_{(it)} + (1 - \text{Y}_{(it)}) \log (1 - \hat{\text{Y}}_{(it)}))$$

$$\text{J} (\text{W}_\text{e},B_S) = \frac{1}{\text{d}} \sum_{j(it)=1}^{d} (L(\text{Y}_{(it)} , \hat{\text{Y}}_{(it)})) \quad (6)$$

Where d = Total data points present in dataset and employed using same loss equation (3).

$$= - \frac{1}{\text{d}}\sum_{j(it)=1}^{d} (\text{Y}_{(it)}\log\hat{\text{Y}}_{(it)} + (1 - \text{Y}_{(it)}) \log (1 - \hat{\text{Y}}_{(it)}))$$

7. Gradient Descent:

The cost function in several machine learning algorithms is lowered using the optimization procedure gradient descent. It is used to update the learning model's parameters.

$$\text{W}_e = \text{W}_{e_1} - \text{L}_\text{r}\text{*d W}_e \quad (7)$$

$$B_{S_2} = B_{S_1} - \text{L}_\text{r}\text{*d } B_S \quad (8)$$

Where, W_e = Weight of the input features
B_S = Bias
L_r = Learning Rate
d (W_e) = Partial derivative of cost function with respect to Weights (W_e)
d (B_S) = Partial derivative of cost function with respect to Bias (B_S)

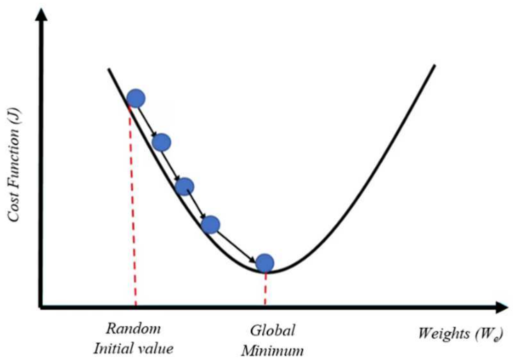


Fig: 4 Gradient Descent for LR

8. Learning Rate:

The learning rate for Gradient Descent needs to be carefully chosen for it to function. The speed at which updating the parameters depends up on the learning rate. If the learning rate is too high, optimal value can be overshoot. Similarly, if it is too small, it will take a long time to reach the optimal values. Therefore, it is important to use a properly calibrated learning rate.

C. K-NEAREST NEIGHBOURS MODEL

One of the simplest supervised machine learning models, K-NN is used for feature-based classification and regression. It measures the distance between two data points using the Euclidean distance formula. The parameter K in the K-NN system denotes the number of closest neighbours to take into account while voting by majority. The accuracy and result are improved by parameter tuning, which involves selecting the appropriate value of k. For greater precision.

1. Choose the value of K:

For better voting, K should be the square root (SQRT) of all the data points and K should be an odd number for proper voting otherwise, voting for the output might get confusing for out model. The distance Z in the formula is known as the Euclidean Distance. In this instance, considering that if K=5. Such that the outcome is decided by majority vote as shown in fig5.

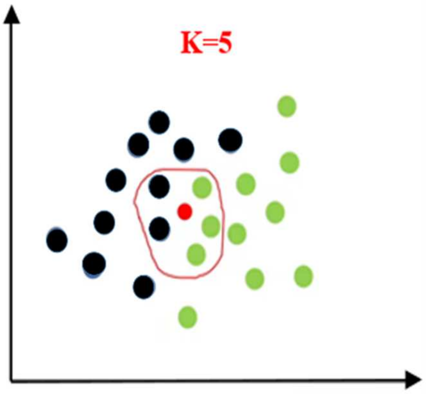


Fig: 5 K-Nearest Neighbour plotting

2. *Calculate the distance:*

Using the Euclidean distance calculation after plotting the data points, model identifies the five samples that are closest to the test sample. The formula below is used to get the Euclidean distance between any two supplied data point locations.

$$Z = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{9}$$

Where, Z=Euclidean distance
x_1, y_1 represent the position A
x_2, y_2 represent the position B

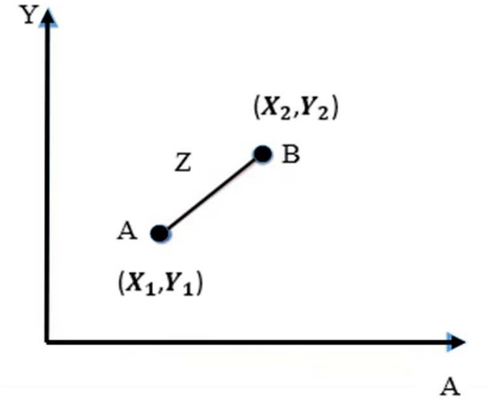


Fig: 6 Euclidean formula (Distance formula)

3. *Sort the nearby points according to Category:*

The graph shows that of the five closest points, three of them have the class label "green," and two have the class label "blue." Therefore, it may be inferred from the majority that red data point (x, y) belongs to a class labelled "green."

4. *Weights and Attributes*

As has been seen, unique x and y dimensions are possessed by every sample point. Features or characteristics, in addition to the coordinates, represent the points in the parameter space. A significant role is played by these characteristics in classifying the points into different categories. Therefore, in addition to locating the closest k-points, the Euclidean distance must be determined using all of those characteristics.

$$x_q = (x_{q1} , x_{q2} , x_{q3} \dots x_{qn})$$

$$y_p = (y_{p1} , y_{p2} , y_{p3} \dots y_{pn})$$

$$\text{Euclidean distance } (x_q, y_p) = \sqrt{\sum (x_q - y_p)^2} \tag{10}$$

It is now known that when predicting the outcome using KNN regression model, the mean values of class labels are taken into account. During classification, the average of the class labels must be used rather than the majority when high values of k are considered. A few of them are listed below.

a) *Noise in attributes:* Due to the existence of noise, it is possible that all features of the test sample may not be captured by the sample point closest to it instead, a sample point that is slightly farther away may be able to do so.

b) *Noise in class labels:* There is a considerable likelihood that test samples will be partially misclassified due to the noise in class labels.

c) *Overlapping class labels:* When class labels are overlapped, the proper class labels cannot be assigned to sample test points by the algorithm.

IV. RESULTS AND DISCUSSION WITH EVALUATION METRICS

Following data processing, the training data set is separated, and two ML classifier algorithms KNN and LR were used in this study. To assess the performance of ML models various assessment criteria including F1 score, recall, precision, and accuracy, were used.

A. *Classification Accuracy:*

When rating categorization models, accuracy is one of the main factor to consider. Accuracy is defined as the proportion of forecasts that were successfully anticipated by the model. The official elaboration of accuracy is given below.

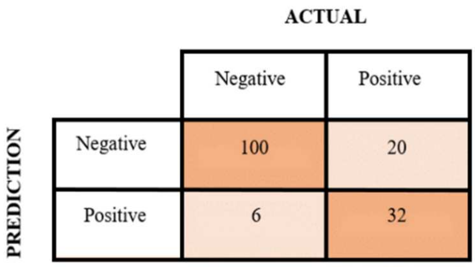
$$\text{Accuracy } (\%) = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \text{ x } 100 \tag{11}$$

TABLE I. ACCURACY OF LR AND K-NN

Models	*Train Accuracy*	*Test Accuracy*	*Test accuracy [2]*	*Test accuracy [12]*
LR	86.05%	83.54%	74%	-
K-NN	82.93%	78.23%	77%	73.33%

B. *Confusion Matrix:*

To demonstrate the effectiveness of a categorization system, a confusion matrix is utilized. In a confusion matrix, the results of a classification model are displayed and recorded. A 2x2 binary classification matrix is employed, with actual results on one axis and expected results on the other. The phrases True Positive (TP$_s$), True Negative (TN$_s$), False Positive (FP$_s$) and False Negative (FN$_s$) in the confusion Matrix are clarified using proposed LR model confusion matrix. It can be shown as follows



		ACTUAL	
		Negative	Positive
PREDICTION	Negative	100	20
	Positive	6	32

Fig: 7 Confusion matrix Representation

True Positive (TP$_s$): The positive class is reliably predicted by the model (both prediction and actual are positive). In the above scenario, the presence of diabetes is correctly predicted in 32 individuals by the model.

True Negative (TN_s): Correct predictions for the negative class are made by the model (both the estimated and actual values are negative). In the above illustration, the presence of diabetes is predicted adversely in 100 individuals by the model.

False Positive (FP_s): The forecasting of a negative class by a model is incorrect (predicted-positive, actual-negative). 6 persons in the aforementioned scenario are flagged as potentially having diabetes despite the fact that they do not.

False Negative (FN_s): When the positive class is incorrectly predicted by the model (predicted negative but actual positive), diabetes are expected to be present in 20 individuals in the earlier instance. The formulas are given by.

$$\mathrm{TP_s\ Rate} = \frac{\mathrm{TP_S}}{\mathrm{Actual\ Postive}} = \frac{\mathrm{TP_S}}{\mathrm{TP_S+FN_S}} \quad (12)$$

$$\mathrm{FN_s\ Rate} = \frac{\mathrm{FN_S}}{\mathrm{Actual\ Postive}} = \frac{\mathrm{FN_S}}{\mathrm{TP_S+FN_S}} \quad (13)$$

$$\mathrm{TN_s\ Rate} = \frac{\mathrm{TN_S}}{\mathrm{Actual\ Negative}} = \frac{\mathrm{TN_S}}{\mathrm{TN_S+FP_S}} \quad (14)$$

$$\mathrm{FP_s\ Rate} = \frac{\mathrm{FP_S}}{\mathrm{Actual\ Negative}} = \frac{\mathrm{FP_S}}{\mathrm{TN_S+FP_S}} \quad (15)$$

The True Positive Rate (TP_s Rate), False Positive Rate (FP_s Rate), True Negative Rate (TN_s Rate), and False Negative Rate (FN_s Rate) can be used to determine whether the model is operating effectively or not despite unbalanced data. To achieve that, FP_s Rate and FN_s Rate should be as low as possible, while TP_s Rate and TN_s Rate should both have high values. Other performance indicators including precision, recall and F_1 score can be calculated using TP_s, TN_s, FN_s, and FP_s.

C. Precision(P_{re}) and Recall(R_e):

Precision: The model's precision measures how well it can forecast a certain category. The fraction of all positively anticipated events that actually occur is represented as precision. Where the value lies in between 0 to 1.

$$\mathrm{Precision}\ (P_{re}) = \frac{\mathrm{TP_S}}{\mathrm{TP_S + FP_S}} \quad (16)$$

Recall: How frequently the model was able to identify a certain category is shown by recall. It is fraction measure of all positives are expected to be positive. It is the same as the TPs rate.

$$\mathrm{Recall}\ (R_e) = \frac{\mathrm{TP_S}}{\mathrm{TP_S + FN_S}} \quad (17)$$

D. F_1 Score:

The harmonic mean of recall and precision is represented by it. Both FPs and FNs are taken into count. As a result, it performs admirably with an uneven dataset.

$$F_1\ \mathrm{Score} = \frac{2}{\frac{1}{Pre}+\frac{1}{Re}} = 2 * \frac{(P_{re} * R_e)}{P_{re} + R_e} \quad (18)$$

TABLE II. PRECISION, RECALL AND F_1 SCORE FOR TEST DATA

SI NO	ML Model	Precision	Recall	F_1 Score
1	LR	0.833	0.9433	0.8848
2	K-NN	0.8081	0.9123	0.852

V. CONCLUSION

The use of ML technology is thought to be beneficial in illness diagnosis. Data pre-processing of the PIDD was carried out in the present study along with the implementation of two types of classification model, to achieve higher prediction in identifying diabetes at early stage. Expression of accuracy have been derived to solve the classification issue. In comparison with previous research works [2, 12], the proposed LR and KNN classification algorithm tuned with hyper parameters have yielded far more superior accuracy. Versatile evaluation metrics were evaluated using the test data set and further trained to achieve better precision, recall and F_1 measure. The outcomes of the proposed strategy thus verify that LR algorithm outperformed the K-NN algorithm. The accuracy of LR and KNN are 83.54% and 78.23% respectively.

REFERENCES

[1] M. K. Hasan, M. M. Hossain, S. S. Islam, M. A. Hossain, and M. A. Karim, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516-76531, 2020.

[2] M. A. Sarwar, F. A. Mahmud, M. T. Rahman, and M. H. Kabir, "Prediction of diabetes using machine learning algorithms in healthcare," in 2018 24th International Conference on Automation and Computing (ICAC), IEEE, 2018, pp. 1-6.

[3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578-1585, 2018.

[4] D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importance for diabetes prediction using machine learning," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, 2018, pp. 1236-1241.

[5] A. M. Posonia, S. Vigneshwari, and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), IEEE, 2020, pp. 569-573.

[6] S. Edeghere and P. English, "Management of type 2 diabetes: now and the future," Clinical Medicine, vol. 19, no. 5, pp. 403, 2019.

[7] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2019, pp. 909-914.

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104-116, 2017.

[9] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Frontiers in Genetics, vol. 9, p. 515, 2018.

[11] T. O. Ayodele, "Types of machine learning algorithms," in New Advances in Machine Learning, vol. 3, pp. 19-48, 2010.

[12] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Computer Science, vol. 216, pp. 21-30, 2023.

[13] K. M. Ghori, R. A. Abbasi, M. Awais, M. Imran, A. Ullah, and L. Szathmary, "Performance analysis of different types of machine learning classifiers for non-technical loss detection," IEEE Access, vol. 8, pp. 16033-16048, 2019.