RESEARCH ARTICLE OPEN ACCESS

A Machine Learning Approach to Predict Diabetes Disease

K.Surya Dev, N.Srushtika, P.Prathap Reddy, PremTeja*, Mrs.R. Durga Devi**

*UG Student, Department of Information Technology, Malla Reddy Engineering College, Hyderabad, India.
**Assistant Professor, Department of Information Technology, Malla Reddy Engineering College, Hyderabad, India.

Abstract:
This paper deals with the prediction of Diabetes.Disease by performing an analysis of five supervised machine learning algorithms, i.e. K-Nearest Neighbors, Naive Bayes, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset , we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

Keywords–Diabetes, Disease, Machine Learning (ML), Disease Risk Analysis, Confusion Matrix, Body Mass Index (BMI), Precision, Recall, F1-Score

I. INTRODUCTION

In this day and age, one of the most notorious diseases to have taken the world by storm is Diabetes, which is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin. Due to the many criterion to be taken into consideration for an individual to harbour this disease, it's detection and prediction might be tedious or sometimes inconclusive. Nevertheless, it isn't impossible to detect it, even at an early stage.

Diabetes is increasing day by day in the world because of environmental, genetic factors. The numbers are rising rapidly due to several factors which includes unhealthy foods, physical inactivity and many more. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels, Blood Pressure, etc., play an important role to the contribution of the disease.we can see that the number of cases is rising every year and there is not slowing down in the active cases. It is a very crucial thing to worry as diabetes has become one of the most dangerous and fastest diseases to take the lives of many individuals around the globe. Machine Learning is very popular these days as it is used everywhere, where a large amount of data is present, and we need some knowledge from it. Generally, we can categorize the Machine Learning algorithms in two types but not limited to-

With the rise of Machine Learning and its relative algorithms, it has come to light that the significant problems and hindrances in its detection faced earlier, can now be eased with much simplicity, yet, giving a detailed and accurate outcome. As of the modern-day, it is comprehended that Machine Learning has become even more effective and helpful in collaboration with the domain of Medicine. Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual. Such early tries can lead to the inhibition of disease as well as obstruction of permitting the disease to reach a critical degree. The work which will be described in this paper is to perform the diabetes disease prediction using machine learning algorithms for early care of an individual.

II. LITERATURE SURVEY

In this, they have used the WEKA tool for data analytics- for diabetes disease prediction on Big Data of healthcare. They used the publicly available data-set from UCI and applied different machine learning classifiers on it. The classifiers which they incorporated are Naive Bayes, Support Vector Machine, Random Forest and Simple CART. Their approach starts with accessing the dataset, preprocess it Weka tool and then did the 70:30 train and test split for applying different machine learning algorithms. They did not go with the cross-validation step as it is imperative to get the optimal result and accurate results as well.

The authors in this also used the publicly available dataset named as Pima Indians Diabetes Database for performing their experiment. Their framework of performing the prediction starts with the dataset selection and then with data pre-processing. Once the data was pre-processed, they applied three classification algorithms, i.e., naive Bayes, SVM and Decision tree. As they incorporated different evaluation metrics, they did compare the different performance measure and comparatively analyzed the accuracy. The highest accuracy achieved with their experiment was 76.30%. Like they have also not practised Cross-validation.

In this, the authors proposed the neural network-based diabetes disease prediction on Indians Pima Diabetes Dataset. They have used several hidden layers to find patterns in the data, and with the help of those patterns, they predicted the outcome. They name their proposed algorithms as ADAP, which is a custom neural network with multiple partitions and with the set of association weights and units. They managed to achieve a crossover point for sensitivity, and specificity at 0.76 and are trying to precise their result in future.

The authors in this used a diverse genera of machine learning algorithms like support vector machine, random forest, logistic regression, Decision tree and many more and various types of disease dataset to show the applicability of Machine Learning in disease prediction and analysis. They also accompanied the traditional way of conducting the analysis by using the data pre-processing, feature extraction and selection, classifiers training and testing for producing the end results. They used feature selection for reducing the computational expenses. Also, to get the most optimal outcome, they divided every dataset into 90% training set and remaining 10% testing set. Along with the accuracy measure, they did the cross-validation for every algorithm and showed different results based on different k values for k-fold cross-validation.

III. PROPOSED SYSTEM

To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database. This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old. Our proposed model exists of 5 phases which are shown in the Fig-1.

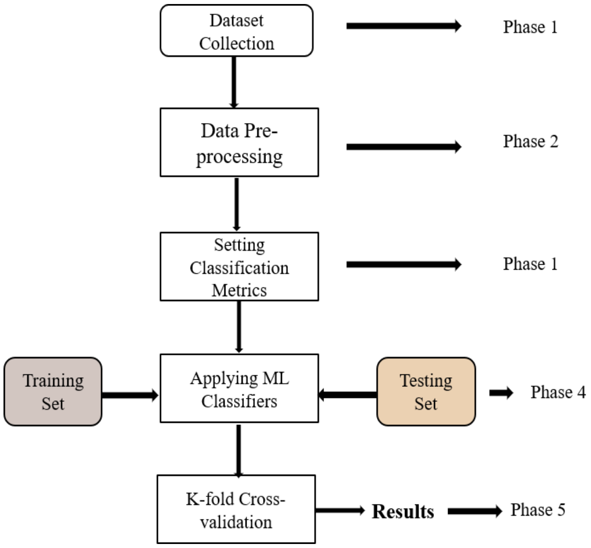

Fig-1: Proposed System Architecture.

A. Data Collection

The dataset mentioned above has eight features which are defined in Table below.

TABLE I
Features and description

Features	Description
Pregnancies	Number of Pregnancies patients had earlier.
Glucose	Glucose level present in the patient.
Blood Pressure	Recorded blood pressure level at that particular time.
Skin Thickness	Skin thickness level of the patient.
Insulin	Amount of Insulin present in the body.
BMI	Body Mass Index of the individual.
Diabetes Pedigree Function	Family history of Diabetes disease.
Age	Age of an individual.

Table-1:

Along with the feature, the dataset also has two label (0-No and Yes). which is the outcome of the diabetes disease. The detailed information about each attribute or features is discussed below.

Pregnancies: Those who develop gestational diabetes are at higher risk of developing type 2 diabetes later in life. The subjects with a greater number of pregnancies have a higher risk of developing diabetes.

Glucose: The subjects were given an oral glucose test, whereby, they were administered glucose and a reading of their plasma glucose concentration was taken after 2 hours. The subjects with higher levels of glucose concentration after 2 hours have a higher risk of developing diabetes.

Blood pressure: Having blood pressure over 140/90mmHg of Mercury are linked to having increased risk of developing diabetes. Although, certain sub

Skin Thickness: Skin thickness is primarily determined by collagen content and is increased in the case of insulin dependent diabetic patients. The subjectstricep skin fold were measured and results showed that having a skin thickness of 30mm of greater are at a higher risk.

Insulin: Normal insulin levels after 2 hours of glucose administration is 16-166mlU/L. Subjects having lower or higher levels than said value are at a higher risk.

Body Mass Index (BMI): Subjects having a BMI over 25 have a relatively high risk in having diabetes.

Diabetes Pedigree Function: The diabetes pedigree function provides “a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject.” The higher the DPF, the more likely it is for a subject to be diabetic.

Age: Diabetes is prevalent in any age group, but is commonly found in middle aged adults (45 onwards). Taking that into consideration, subjects with in the
higher age group have a higher expectancy of diabetes.

B. Data Pre-processing

The dataset, which is quoted above, has lapsed and have shed data. To make the dataset serviceable and obtain the knowledge from it, we have performed data preprocessing. In order to handle erroneous data, we have analyzed the dataset for the unusual entries and fixed them manually. Missing values are handled with the help of calculating the standard deviation of that particular feature and allotting it to the missing spaces. To make the dataset useful, we have used Pandas and NumPy library for handling the dataset efficiently and easy data handling throughout the experiment.

C. Setting Classification Metrics

To classify disease and get a prediction result, we need to set a few metrics which will help us in predicting the Diabetes disease. Since we are using scikit-learn (Sklearn) machine learning library [8] for our experiment, we have used confusion matrix as the classification measure metrics. All the used metrics, i.e., *Precision, Recall, F1-Score and Accuracy* in our analysis, are listed below.

• **Precision** (***P***)is defined as the number of true positives (*Tp*) over the number of true positives plus the number of false positives (*Fp*). Mathematically,

$$P = Tp / Tp + Fp \quad \text{----(1)}$$

• **Recall** (***R***)is defined as the number of true positives (*Tp*) over the number of true positives plus the number of false negatives (*Fn*).

$$R = Tp / Tp + Fn \quad \text{-----(2)}$$

• **F1-Score** (***F*1**) is defined as the harmonic mean of precision and recall.

$$F1 = (2 * P * R) / P + R \quad \text{-----(3)}$$

• **Accuracy** (***A***)is defined as follows.

$$A = (Tp + Tn) / (Tp + Tn + Fp + F \quad \text{-----(4)}$$

D. Applying Machine Learning Algorithms

For our experiment, we will perform 5 supervised machine algorithms on the pre-processed dataset. The algorithms which we used are as follows

1) K-Nearest Neighbor (KNN) with K=10
2) Naive Bayes (NB)
3) Decision Tree (DT)
4) Random Forest (RF)
5) Support Vector Machine (SVM)

1. Nearest Neighbor

In KNN, "nearest" specifies a distance measure. KNN regression is used to estimate the average of the geometric targets of the K-Nearest Neighbours and it also used to find an opposite distance weighted average of the K-neighbour [18]. It uses the same distance functions to calculate distance between two data points what KNN classification technique is used. The distance function used in K-NN regression is

$$d=\sqrt{(x2-x1)^2+(y2-y1)^2}$$

2. Naive Bayes (NB)

Naive Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

The formula for Bayes' theorem is given as:

$$\mathbf{P(A/B) = P(B/A)\ P(A) / P(B)}$$

3. Decision Tree Regression

In this regression, the predictor space is dividing into the set of possible values for X1, X2 . . ., Xp into j separate and non-overlapped areas, A1, A2 . . ., Aj. For each observation that falls into the area Aj, produces the similar prediction [19], which is simply the mean of the response values for the training observations in Aj. Here the main objective is to locate areas A1, A2 . . ., Aj to minimize the RSS by,

$$\mathbf{RSS = \sum Jj=1 \sum i\in Aj(yi - y\hat{}\ Aj)\ 2} \quad \text{----- (9)}$$

Where, the training values inside the j-th area are represented by their mean response.

4. Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions and it predicts the final output. The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

5. Support Vector Machine

Support Vector Machine is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the

SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane.

6. K-Fold Cross Validation

Lastly, we have used another evaluation method, namely K- Fold cross-validation for precise use of dataset and also for calculating most optimal accuracy results. In K- fold cross-validation the dataset is partitioned into K- different folds (in our case K= 10) and then in each cycle one of the folds (say Fold-1) is examined with the remaining (k-1) folds. This process is going to recur continuously till all the folds get examined. For a better understanding of K- Fold Cross-validation, we have pictorial representation of this evaluation metrics in Fig.

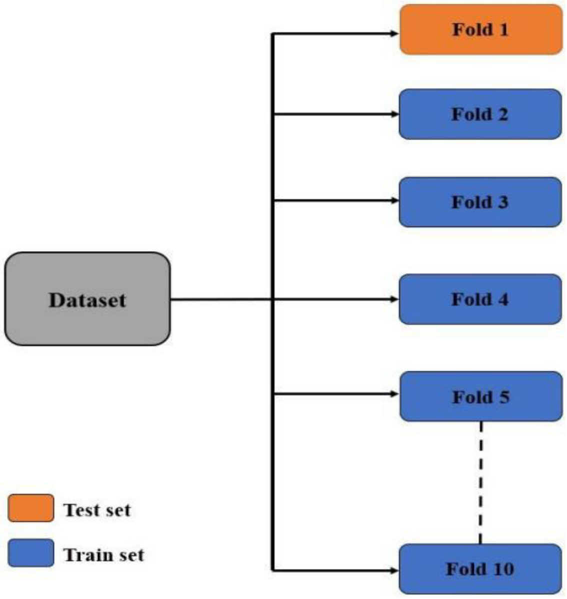

Fig-2: Different cross folds of KNN

IV. EXPERIMENTAL RESULTS

To perform our experiment, we have split the dataset into training and testing set with the ratio of 80:20, respectively. We used python 3.6 version for applying all of our machine learning classifiers with the system configuration of 8 GB RAM and Intel i5 9th Generation mobile processor. We can see in the Fig that KNN classifiers achieved the highest accuracy of 76% after performing 10- cross-validation and the other classifiers also managed to achieve good accuracies of above 70%. For the detailed information of each metrics which we have incorporated in our experiment is given in Table II.

TABLE II
Performance Metrics of different Classifiers Models

Classifier	P	R	F1	A (10-fold)
KNN	0.76	0.73	0.75	0.76
SVM	0.73	0.74	0.73	0.75
NB	0.74	0.74	0.74	0.74
DT	0.72	0.71	0.71	0.71
RF	0.70	0.71	0.71	0.71

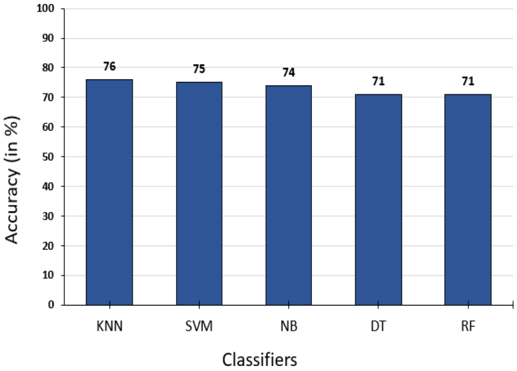

Fig-3: Accuracies of different classifiers after 10-fold cross-validation.

Then we made an accuracy comparison between the training and testing accuracy. These accuracies are useful for knowing whether our model is facing the problem of overfitting or underfitting. Higher Training accuracy implies that our model is suffering from model overfitting, and the respective classifiers models are learning less from the training data and more from the test data. On the other hand, if the testing accuracy is higher, it implies that our classifier is suffering from model underfitting and the models are learning significantly fewer rules from the training data, and they are not able to build relationships with test data. For a good result, we need comparable training and testing accuracies.

We saw training and testing accuracy result in Fig.3 where all the training accuracies were more significant than testing accuracies which tell us that our classifiers suffer from model overfitting and they are learning more rules from testing data. To get higher accuracy, we need more instances (patient data) in the dataset, which can draw relationships between training and testing data to omit model overfitting and achieve better performance in terms of accuracy and precision. Also, we can see that the KNN has least training and testing accuracy difference which tell us that it suffers significantly less form model overfitting and it will be the most suitable classifier to predict the diabetes disease as far as this dataset is concerned. The SVM and RF classifiers have a substantial amount of training and testing accuracy difference, and hence it suffers from the model overfitting among all the machine learning classifiers.

Finally, we checked the computational timings of all the classifiers. In order to calculate the training and testing time, we used the python *time* library. The time needed for training (learning from the given data) the specific model or machine learning algorithm is known as the training time. On the other hand, the time needed for testing (checking the results by cross verifying the new data to trained data) is called as testing time. Computational time (C_t) is estimated with the following formulas below-

$$C_t = T_t + T_s$$

where $T_t$and Ts represents training time and testing time respectively.

V. CONCLUSION

One of the significant impediments with the progression of technology and medicine is the early detection of a disease, which is in this case, diabetes. However, in this study, systematic efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 76% using the K-Nearest Neighbors classifiers. With this being said, it is hopeful that we can implement this model into a system to predict other deadly diseases as well. There can be room for further improvement for the automation of the analysis of diabetes or any other disease in the future. In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. We will be incorporating more Machine Learning and Deep learning models for achieving better results as well.

REFERENCES

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045:

[2] Results from the international diabetes federation diabetes atlas, 9th edition,” Diabetes Research and Clinical Practice, vol. 157, p.107843, 2019.

[3] A. Mir and S. N. Dhage, “Diabetes disease prediction using machine learning on big data of healthcare,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6.

[4] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia Computer Science, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918308548

[5] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, “Using the adap learning algorithm to forcast the onset of diabetes mellitus,” Proceedings - Annual Symposium on Computer Applications in Medical Care, vol. 10, 11 1988.

[6] P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

[7] Wes McKinney, “Data Structures for Statistical Computing in Python,” in Proceedings of the 9th Python in Science Conference, Stefan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.

[8] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen,Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H.vanKerkwijk, M. Brett, A. Haldane, J. F. del R’ıo, M. Wiebe, P. Peterson, P. G’erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Edouard Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, p. 28252830, 2011.

[10] S. Yadav and S. Shukla, “Analysis of k-fold cross-validation over hold out validation on colossal datasets for quality classification,” in 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78–83