Prediction of Diabetes Using Machine Learning: Analysis of 70,000 Clinical Database Patient Record

Sony M Kuriakose[1], Peeta Basa Pati[2], Tripty Singh[3]
Department of Computer Science and Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
[1]BL.EN.R4CSE21010@bl.students.amrita.edu, [2]0000-0003-2376-4591, [3]0000-0002-3688-4392

2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-6654-5262-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCNT54827.2022.9984264

Abstract—
Diabetes is a disease that occurs when the blood glucose level or blood sugar level meets high values. The level of sugar is increased in the human body is due to several factors like obesity, physical inactivity, gender, age, family history, food habits, and so on. Based on these attributes and with the help of machine learning techniques one can foresee diabetes. According to the increasing morbidity, the number of patients who suffer from diabetes will reach 642 million in 2040, which indicates one of the ten adults in the world will suffer from diabetes. Algorithms that are used in Machine learning can apply in the various medical health field to detect and predict diseases. In this paper, we applied Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) machine learning algorithms to predict diabetes in Diabetes 130-US hospitals for the years 1999-2008 Data Set and Pima Indian Diabetes Dataset. We made a comparative study of the accuracy of all machine learning algorithms. In our diabetic prediction model, we got a higher accuracy value for the random forest algorithm.

Keywords- Diabetes Prediction, Random Forest Algorithm

I. INTRODUCTION

The rapid growth of population in world and health maintenance is moving to a crucial subject in world wide. Many diseases are causing more problems in recent years. By using machine learning technologies in healthcare leads early detection and prevention of many diseases. Many machine learning algorithms can predict and diagnose disease like cancer, diabetes with high accuracy. Diabetes is a type of lethal disease that sufficient to cause death. It leads to various disorders like heart failure, blindness, urinary organ diseases. High value of computational time and low value of accuracy in prediction is the main disadvantage of the existing model[13]. To avoid this problem, they have proposed a model that predict and classify the presence of diabetes disease in e- health care environment. For implementation we took two machine learning algorithms one is Logistic Regression and another is Random Forest.

Diabetes is a disease where the level of blood sugar or glucose level of blood touches a higher value. Type 1, Type 2 and Gestational diabetes are different types of diabetes. Less amount of insulin production in the human body leads to Type1 diabetes and Type2 occurs to insulin resistant blood cells and Diabetes occur during pregnancy is called Gestational Diabetes[12]. Diabetes causes different health issues like strokes, eye issues, nervous related problems and also damages in kidney. The key cause of diabetes is not clear but researchers found that diabetes diseases play a vital role in a human genetic factor. Diabetes is untreatable disease. Diabetes disease can be treated by keep up the levels through good medicines and treatment. Detection and prediction of diabetes in first stage of disease, leads to avoid the damages of organs caused by diabetes.

Early prediction of diabetes can be avoiding the death rates and rescue the life of human beings. Our work focusses on selecting important features that cause diabetes. In our work, Pima Indian Diabetes Dataset is used and we applied different algorithms used in machine learning to predict disease. Machine Learning Techniques are used to predict diabetes. By using these supervised machine learning algorithms, we can train computing devices to learn without being explicitly programmed. By building various classification models from collected dataset, Machine Learning Techniques provide efficient result. All the collected information's can be used to predict diabetes. Different types of Machine Learning algorithms are used to do prediction, but it's difficult to find good technique. For finding which algorithm fits for diabetes prediction in our dataset, we used logistic Regression, SVM, Random Forest and KNN for prediction. We used PIMA diabetes dataset and Impact of HbA1c Measurement on Hospital Readmission Rates that Analyze 70,000 Clinical Database Patient Records[11].

II. RELATED WORK

Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang [1] proposed a model that used machine learning techniques and Predict Diabetes Mellitus. They used neural network and also algorithms like Decision tree and Random forest classifier to predict diabetes. Proposed system used dataset from one of the hospitals in Luzhou, china. According

to their studies they got prediction accuracy of 80% by using random forest that included all attributes. In this paper Value of dimensionality is decreased by Principal Component analysis and mRMR that is minimum redundancy maximum relevance). The accuracy rate of prediction when using PCA is not good, and mRMR produced best result[14].

Aishwarya Mujumdara, Dr. Vaidehi Vb [2] proposed a model to predict diabetes. They used Machine Learning Algorithms for Predicting diabetes. They created a prediction model for diabetes. This model is used for better classification of diabetes and also included some external factors that leads to diabetes like level of glucose, level of BMI, Age of the patient , Insulin value of the patient , etc that are regular factors. The accuracy of Classification is boosted by comparing new dataset existing dataset.

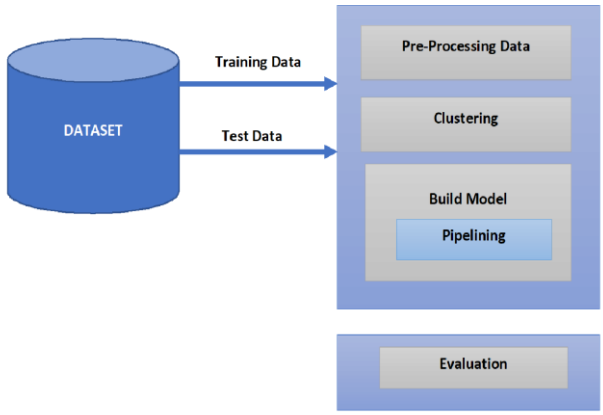

Fig 1: Diabetes Prediction Model

Fig 1: shows the Architecture of diabetes prediction model. Architecture is divided into five parts:
1. Collection of Dataset
2. Pre-processing of data: It normalize the data in the dataset.
3. Clustering: Each person details in the data set is divided into two classes one is diabetic and other one is non diabetic by using K-means clustering
4. Build Model
In this model they used two algorithm Algorithm1 is used for the prediction of diabetes. Predicted diabetes by using various machine learning algorithms. Machine learning techniques includes, the main supervised algorithms like Random Forest Classifier, Gradient Boost algorithm, Decision Tree algorithms(CHAID, CART, C4.5) SVM(Support Vector Machine), Extra Tree algorithm and also Ada Boost algorithm, Logistic Regression, Linear Discriminant Analysis, KNN, Naive Bayes algorithm and Bagging algorithm[5]. Second technique is used to predict diabetes using pipeline. It create pipeline for algorithms giving highest accuracy. The predicted results of model is evaluated by using various evaluation metrics like classification accuracy, confusion matrix and f1-score.

Umair Muneer Butt , Sukumar Letchmunan , Mubashir Ali , Fadratul Hafinaz Hassan ,Anees Baqir , and Hafiz Husnain Raza Sherazi proposed[6] a model for classifying Diabetes Classification and Predicting diabetes for Healthcare Applications[3]. Random forest and Multilayer perceptron are the two random forest classifiers used in the model. Long Short-Term Memory, Moving Averages, Linear Regression (LR) are used for predictive analysis. PIMA Diabetes dataset is used as benchmark for the experimental evaluation of the model. Observations of this analysis are MLP gives accuracy of 86.08% and upgrade the accuracy of diabetes prediction of 87.26% by using LSTM.

Convolutional LSTM Networks is a productive approach for Detecting Diabetes using techniques in Deep Learning and it is proposed by Mitushi Soni and Dr. Sunita Varma[4]. CLSTM(Convolutional Long Short-term Memory) is used in the model and they proposed a model for classifying and predicting diabetes and contrast it with the existing Pima Indians Diabetes Database (PIDD) architecture . They analyzed the results of various techniques in classification. These techniques are improved by an efficient method for data preprocessing and is called as multivariate imputation. Multivariate imputation is implemented by using chained equations. We can train the available data of patients and from that we can diagnose the patients are diabetic or not by using classifier with higher accuracy on the dataset.

III. PROPOSED METHODOLOGY

Objective of the paper is prediction of diabetes in two diferrent dataset by using machine learning algorithm with higher value of accuracy .

Phases of Proposed Methodology:

1. Dataset Description- We used two data sets
i) Pima Indian Diabetes Dataset: Pima Indian Diabetes Dataset is used to collect the data. The data is collected from UCI repository it is called Pima Indian Dataset for Diabetes. There are seven attributes in the dataset of 768 patients. Class variable for for each data points is attribute Number 9. This class variable outcome indicate 0 for positive diabetes and 1 for negative diabetes.

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness Age
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

Fig 3.1: Pima Indian Dataset attributes for diabetes

ii) Dataset used is 130 United states hospitals patient details for years 1999 to 2008: The dataset includes 130 US hospitals ten years 1999 to 2008 of clinical care informations. It contains over 50 features representing details of patient and hospital outcomes. Database included 70,000 analysis of Clinical Database Patient Records. Large clinical database of 74 million encounters that are unique and equivalent to seventeen million patients and no repeated patient details) to detect future path which improves the safety of patient. Inpatient diabetes around 70,000 s were identified with adequate detail for analysis.

2. Data Preprocessing:

Most important process is Data preprocessing. Data related to mostly includes missing value and also includes impurities that leads the effectiveness of data. Quality and effectiveness can be increased after mining step, Preprocessing of data is done after mining process to improve the quality and effectiveness of data. Mining process plays vital role for giving accurate result and prediction when we are using Machine Learning Techniques on the dataset. . Splitting of data- Data is splitted into train and test data, training data set is trained using algorithm and have testing data set aside. Training model is create after the training process and its created using logic and algorithms and also the information getting from the attributes in the training dataset. Bring all the attributes under same scale is the basic aim of normalization.

3. Apply Machine Learning:

1. Logistic Regression- Logistic regression[8] is used to calculate the probability of outcome, based on one or more predictors of a binary response. It will produce continuous or discrete values. When we want to analyze or differentiate some data items into different classes we use Logistic Regression. It analyse and predict a patient is dibetic or not and that data is converted to binary form, that means in zeros and ones, 0 represent positive and 1 represent negative for diabetes. Target and predictor variable relationship is defined by logistic regression algorithm. Probability of positive and negative class is calculated by using sigmoid function in Logistic regression model.

$$\textbf{Sigmoid Function Z} = \frac{1}{1+e^{-(c+dx)}}$$

Here z is probability,
c and d are parameter of Model.

2. K-Nearest Neighbor- Machine learning technique that comes under supervised machine learning algorithm. KNN consider that closed values are near to each other. KNN predicts the new data point by finding the neareast value of data in the training data set. Euclidean distance gives the closeness. Two points P and Q Euclidean distance is defined by the following equation:-

$$\textbf{D(P,Q)} = \sum_{i=1}^{n} (\boldsymbol{Pi - Qi})^2 .$$

Using similarity KNN create new work . KNN algorithm stores all the informations based on the similarities and then analyse it. Distance between the points are finding by using tree like structure. N is equal to the number of nearby neighbors and value of N is always a positive integer. From set of classes Neighbor's value will choose.

Algorithm:

- Take a sample dataset here it is Pima diabetes data set .
- Select features and rows for test data.
- Calculate the Euclidean Distance

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

- Choose a value of N. is the number of nearest neighbors(It should be a random value).
- For finding the nth column apply minimum distance and Euclidean distance.
- Calculate the above figures for Output.

3. Support Vector Machine is a supervised machine learning algorithm(SVM). Two classes are separated by Hyperplane that created by the SVM. It will find a hyperplane or collection of a hyperplane in high dimensional space.

Algorithm:

- Find the hyperplane that separate the class better.
- Find the distance from the planes to the data, that distance is called Margin.
- The chance of miss conception is higher when the distance between the classes is low and vice versa.
- Select the highest margin classes
  Margin is the sum of the distance to the positive point and negative point distance.

4. Random Forest: Random Forest is an ensemble machine learning algorithm[9]. It develops decision trees on various samples for classification itusese their majority vote or in the case of regression, it takes average. By using Gin-Index Cost function random forest calculates the best split.

$$\textbf{Gini} = \sum_{k=1}^{n} \boldsymbol{p_x} * (\mathbf{1} - \boldsymbol{p_x})$$

Where k= Each class and P = Proportion of training instances.

In our proposed model we applied logistic regression, SVM, KNN, and random Forest in the Health Facts database with 70000 patients' details and Pima Indian diabetes Database with 768patients' information. We predicted diabetes by using logistic regression in Pima Indian diabetes by using all the features of the n dataset. We also predict diabetes by using both

the database and unchecking the accuracy difference. In both databases, the insulin level is the common feature by using insulin level we predicted whether a patient has diabetes or not.

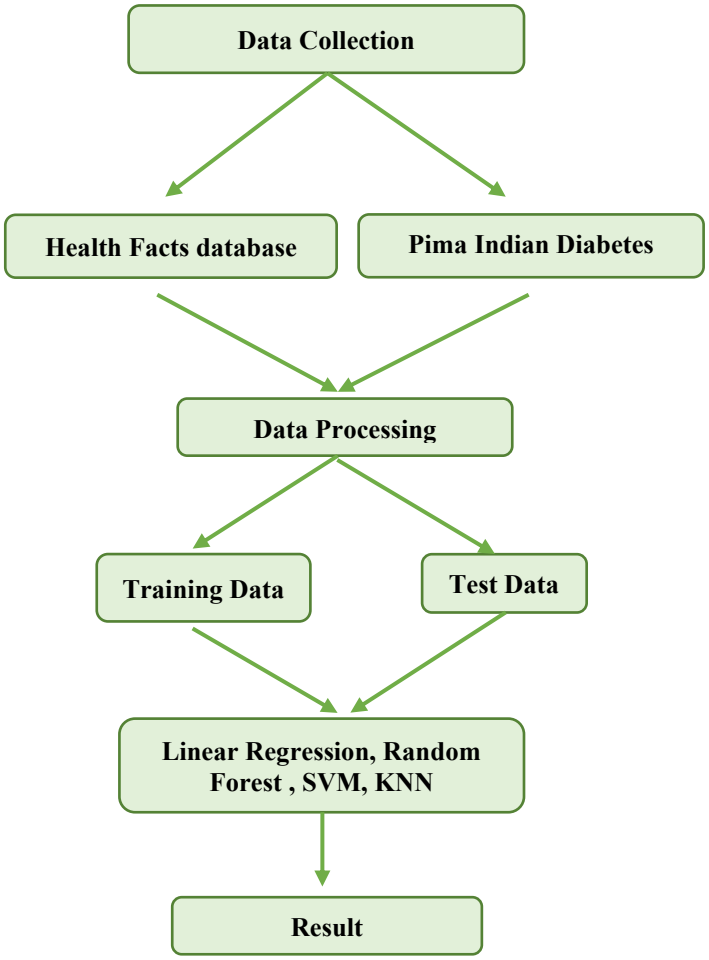

Fig 3: Proposed System flow Chart

IV. RESULT

In our model logistic regression is applied in Pima Data Indian Diabetes Database with 768 patients with all the features and got an accuracy of 73%. The same logistic Regression is applied in Health Facts database with 70000 patients' insulin data and got an accuracy of 77%. Then we applied logistic regression in the Pima Data Indian Diabetes database with 768 patients' insulin data and got an accuracy of 63%.

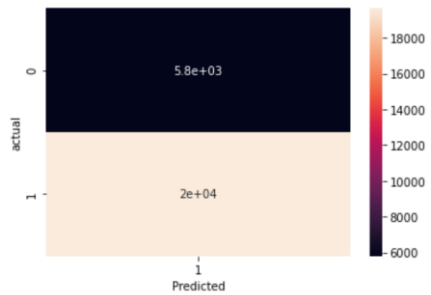

Fig4.1 Accuracy of diabetes prediction using Logistic Regression using Health Facts database with 70000 patients insulin data

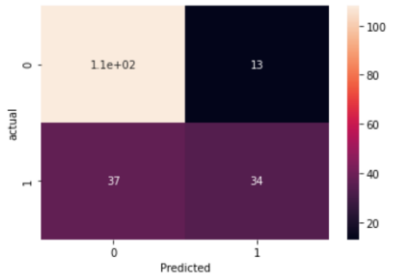

Fig 4.2 Accuracy of diabetes prediction using Logistic Regression in PIMA Indian database with 768 patients with all the features.

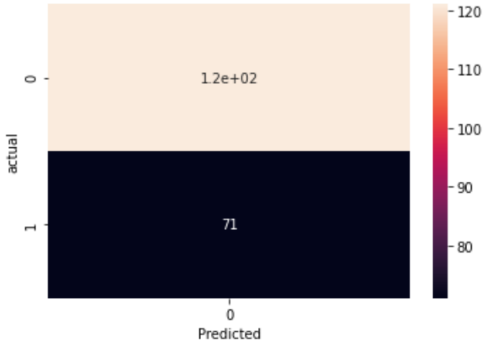

Fig 4.3 Accuracy of diabetes prediction using Logistic Regression in PIMA Indian database with 768 patients insulin data.

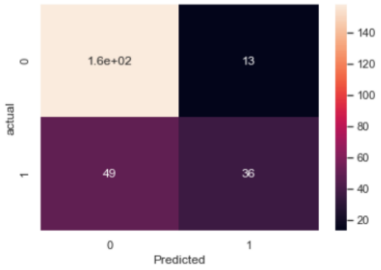

Fig4.4 Accuracy of diabetes prediction using Random Forest Classifier using Health Facts database with 70000 patient's insulin data

Results	
KNN applied in Health Facts database with 70000 patients insulin data	**Accuracy:69%**
Logistic Regression applied in Health Facts database with 70000 patients insulin data	**Accuracy:76%**
SVM applied in Health Facts database with 70000 patients insulin data	**Accuracy:77%**
Random Forest applied in Health Facts database with 70000 patients insulin data	**Accuracy:79%**

Fig 4.5 Accuracy Result of Machine Learning Methods - Health Facts database with 70000 patients

Results	
KNN applied in Pima Data Indian Diabetes Database with 768 patients with all the features	Accuracy:73%
Logistic Regression applied in Pima Data Indian Diabetes Database with 768 patients with all the features	Accuracy:73%
SVM applied in Health Facts database with 70000 patients insulin data	Accuracy:77%
Random Forest applied in Health Facts database with 70000 patients insulin data	Accuracy:80%

Fig 4.6 Accuracy Result of Machine Learning Methods - Pima Data Indian Diabetes Database with 768 patients

Pima Data Indian Diabetes Database	Health Facts database
768 patients' details included in database	70000 patients' insulin data
KNN applied in Pima Data Indian Diabetes Database with 768 patients with all the features and gives accuracy of 73%	KNN applied in Health Facts database with 70000 patients' insulin data gives accuracy of 69%

Fig 4.7 Comparison between Pima and Health facts database

V. CONCLUSION

This paper aims to predict diabetes using machine learning and compare the accuracy of four machine learning algorithms when used in two different datasets. The proposed approach enhances the prediction of diabetes using various Machine learning techniques like SVM, Random Forest classifier, Logistic Regression, and K-Nearest Neighbour (KNN). In the future, by using a machine-learning algorithm we will predict diabetes, which will help prediction of diabetes at its early stages and that helps to reducing the risk of various diseases.

VI. Future Enhancement

Limitation of our propsed work is when we used health facts database we got accuracy of 69% because of more missing values in dataset. We will find new algorithms in future that fix this issue with high prediction accuracy.

REFERENCES

[1] Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques.

[2]World Health Organization, Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World, World Health Organization, Geneva, Switzerland, 2019.

[3]R. Williams, S. Karuranga, B. Malanda et al., "Global and regional estimates and projections of diabetes-related health expenditure: results from the international diabetes federation diabetes atlas," Diabetes Research and Clinical Practice, vol. 162, Article ID 108072, 2020.

[4]M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," 2018 24th International Conference on Automation and Computing (ICAC), 2018, pp. 1-6

[5]P. Moksha Sri Sai, G. Anuradha, VVNV Phani kumar, "Survey on Type 2 Diabetes Prediction Using Machine Learning", Computing Methodologies and Communication (ICCMC) 2020 Fourth International Conference on, pp. 770-775, 2020.

[6]Gaurav Tripathi, Rakesh Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning", Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) 2020 8th International Conference on, pp. 1009-1014, 2020.

[7] Naz and Ahuja, H. Naz, S. AhujaDeep learning approach for diabetes prediction using PIMA Indian dataset
Journal of Diabetes & Metabolic Disorders, 19 (1) (2020), pp. 391-403

[8] Parashar ,A. Parashar, K. Burse, K. RawatA Comparative approach for Pima Indians diabetes diagnosis using lda-support vector machine and feed-forward neural network International Journal of Advanced Research in Computer Science and Software Engineering, 4 (11) (2014), pp. 378-383

[9] A. Marcano-Cedeño, J. Torres, D. Andina A prediction model to diabetes using artificial metaplasticity Proceedings of International Work-Conference on the Interplay Between Natural and Artificial Computation (2011), pp. 418-425

[10] Bala Manoj Kumar P, Srinivasa Perumal R, Nadesh R K, Arivuselvan K, Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier, International Journal of Cognitive Computing in Engineering, Volume 1, 2020, Pages 55-61,ISSN2666-3074,

[11] Y. Tarakaram, Y. Mounika, Y. Lakshmi Prasanna and T. Singh, "Codon Optimization and Converting DNA Sequence into Protein Sequence using Deep Neural Networks," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5.

[12] A. Tripathi, T. Singh and R. R. Nair, "Optimal Pneumonia detection using Convolutional Neural Networks from X-ray Images," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6.

[13] S. Khan, S. Gupta, D. Gupta and S. K. Jha, "A Study on Binary File Conversion using python-based GUI Application," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 451-456.

[14] A. Srivastava, S. Saini and D. Gupta, "Comparison of Various Machine Learning Techniques and Its Uses in Different Fields," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 81-86.