Diabetes Detection With Feature Engineering Based Machine Learning Model

Swapnshri Patel[1], Anjali Singh[2]
[1]Dept of CSE
[2]Professor, Dept of CSE
[1, 2] Aditya College of Technology and Sciences, Satna, M.P

Abstract- *Diabetes can lead to other serious issues such as cardiovascular diseases, renal failure, and neuropathy and contribute to rising mortality rates in diabetic patients. Considering the situation, accurately predicting mortality risks in diabetic patients is crucial for several reasons like identifying high-risk individuals, Studies conducted by WHO and CDC indicate that risk of mortality is very high in diabetic patients and it's hard to predict the insulin amount required for each patient and it gets progressively harder when the patient has additional complications and comorbidities like HIV/Depression/Alcohol Abuse etc. Many influencing factors that could affect the diabetic complications need to be considered and hence there is a need to develop an all-cause mortality prediction model that could be utilized by health- practitioners for devising better diabetic treatment plans, identifying sensitive individuals and controlling the mortality rate. The paper focuses on Type 2 Diabetes Mellitus which generally occurs when insulin is not effectively used by the body due to excess body weight and physical inactivity. The study tracks the mortality status of patients at both the 5- year and 10-year intervals. The work however will not cover the patients with Type 1 Diabetes or gestational diabetes and usage of external datasets for the validation of model is also not within the scope of the work.*

Keywords- Diabetes Mellitus, Mortality, features, Machine learning, XGBoost, AUC, Accuracy.

I. INTRODUCTION

Diabetes is a non-communicable disease that affects the control of blood sugar levels in the body. Blood glucose concentration is normally controlled by insulin and glucagon, two hormones secreted by the beta (β) and alpha (α) cells of the pancreas, respectively. The normal release of the two hormones regulates blood sugar in the body within the range of 70 - 180 mg/dl (4.0 - 7.8 mmol/l). Insulin lowers blood sugar, while glucagon increases blood sugar. However, abnormality of these hormones can lead to diabetes. However, there are many types of diabetes, including different types of diabetes such as type 1 diabetes, type 2 diabetes, and gestational diabetes (gdm). Type 1 diabetes is more common in children; while type 2 diabetes is more common in adults and the elderly, gdm is more common in women and is diagnosed during pregnancy. While insulin secretion does not work in type 1 diabetes due to the destruction of pancreatic beta cells, there is a disorder in insulin secretion and function in type 2 diabetes. Gdm is a glucose intolerance first diagnosed during pregnancy; it may be mild, but it is also associated with high blood sugar and high insulin levels during pregnancy. All of these types can cause a lack of blood sugar in the body, which can lead to serious diseases in the body. In other words, when blood sugar rises above normal, this condition is called hyperglycemia. On the other hand, when it decreases and falls below normal, the condition is called hypoglycemia [1]- [5]. Both conditions can have a negative impact on a person's health. For example, high blood sugar can cause chronic problems and lead to kidney disease, retinopathy, diabetes, heart attack and other tissue damage, while hypoglycemia can also be affected. Short term. It can cause kidney disease, retinopathy, heart disease, and heart attack, and other damage can lead to diabetic coma [1], [2]. Diabetes has become an important health problem in today's world due to its prevalence in children and adults. According to [6], [7] , approximately 8.8% of adults worldwide had diabetes in 2015, and this number was approximately 415 million and is expected to reach approximately 642 million in 2040. More than 500,000 children were killed during this period. And nearly 5 million people died. On the other hand, the global economic burden of diabetes was estimated to be approximately 673 billion dollars in 2015, and is expected to reach 802 billion dollars in 2040. Self-monitoring of blood glucose (smbg) using fingertip blood glucose meters is a diabetes treatment method introduced three years ago [8], [9]. In this way, diabetics measure their blood sugar levels using a finger glucometer on the skin of their fingers three to four times a day. The idea is to provide this: to increase insulin resistance. However, this method is laborious and laborious, and can only be understood if insulin estimation is obtained from small smbg samples. In other words, this may cause the blood sugar in the blood to be higher than normal. To overcome this problem, continuous blood glucose monitoring (cgm) has been introduced, which can provide maximum information about changes in blood sugar within a few days, allowing a good treatment decision to be made for people with diabetes. In this way, blood sugar concentration is constantly

monitored thanks to small devices/systems that monitor the glucose level in the blood environment. These systems can be invasive, minimally invasive or noninvasive. Moreover, cgm systems can be divided into two types: retrograde systems and immediate systems [10]. The introduction and availability of new types of cgm devices/machines have brought new opportunities for diabetics to easily manage their diabetes. Most cgm devices today often use a minimally invasive device to calculate and record the patient's current blood sugar every minute by measuring interstitial fluid (isf). These systems/devices have little effect because they damage the skin but not the blood vessels. There are also non-invasive methods to measure blood glucose concentration, such as using electrical current through the skin into blood vessels in the body [11].

This paper present review and feature engineering-based ML models which get higher accuracy.

II. LITRATURE REVIEW

A diverse literature has contributed to the area of diabetes diagnosis and prediction ranging from the development and performance analysis of novel data mining based techniques for diabetes detection, prediction, and classification, to the survey and review studies, as can be seen in [15]. In [16], various data mining techniques for diabetes detection are reviewed and discussed. Similarly, in [17], a systematic review of the application of data mining techniques for diabetes, as well as the corresponding data sets, methods, software, and technologies, is carried out. Based on this review, it is concluded that data mining has a key role and bright research future in the field of glycemic control. Data mining is used to extract valuable information from diabetes data, which ultimately helps diabetic patients in the management of their glycemic control. Likewise, in [18], a survey is conducted on the application of different data mining techniques, including artificial neural network (ANN), for the prediction and classification of diabetes. The survey shows that ANN outperforms the rest of the techniques with 89% of prediction accuracy.

On the other hand, in [19], the performance of four well known methods, namely J48 decision tree (DT) classifier, KNN, random forests algorithm, and support vector machine (SVM), is evaluated in terms of prediction of diabetes using data samples with and without noise from the University of California Irvine (UCI) machine learning data repository [20]. From the comparative analysis of these techniques, it is observed that J48 classifier performs better in the presence of noise in the data with 73.82% accuracy. Whereas in case of noise-free data, the KNN (k=1) and random forests outperform the rest of the two methods with an accuracy of 100%. Furthermore, in [20], with the help of data mining tools such as WEKA, TANAGRA, and MATLAB, a comparative study of nine different techniques is performed in the light of diabetes prediction using Pima Indian diabetes dataset (PIDD) from UCI machine learning repository [20]. According to the performance analysis, the best classifiers in WEKA, TANAGRA, and MATLAB are J48graft, NB and adaptive neuro-fuzzy inference system (ANFIS) with the corresponding accuracies of 81.33%, 100%, and 78.79%, respectively. Likewise, in [21], the comparison and performance evaluation of various data mining techniques are presented. In [22], a study is conducted based on six diabetes intervention models using SVM classification technique. The comparative analysis shows that smoking cessation is the best intervention with high accuracy. Moreover, in [23], a method based on data driven model is proposed for the glucose prediction using a multi-parametric set of free-living data such as food, activity, and CGM data. In this method, the effect of diet, physical activity, and medication on the glucose control is investigated.

III. PROPOSED WORK

Proposed system is shown below:

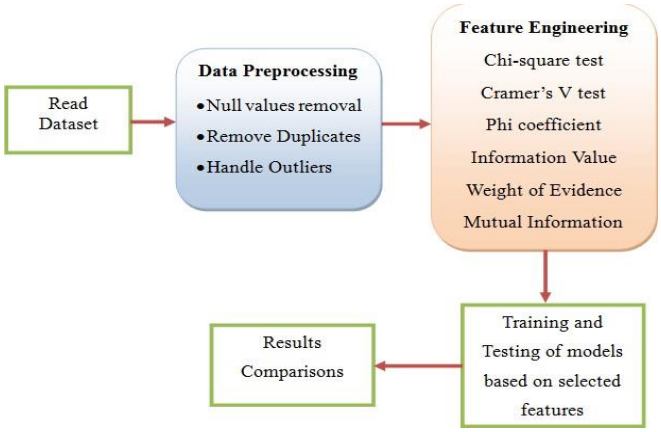

Figure 1: Architecture

Following Feature Engineering steps are applied:

1. Calculate Chi-Square Values: The chi-square test was used to determine the strength of the relationship between the predictor variables and the target variable [63].
2. Cramer's Test:The Cramer's V test was used 9us to gauge the strength of association between pairs of categorical predictors. In Cramer's V test, values range from 0 (indicating no association) to 1 (indicating a perfect association).
3. Phi coefficient:The Phi coefficient measures the strength and direction of association between two

binary variables. In our case, the target variable is mortality, and each predictor is a binary variable indicating the presence or absence of a specific condition or characteristic.

After the EDA and the feature engineering is completed, training and testing will start.

- Put all columns in X except the target column named "mortality".
- Divide the data into train set and test set with size=0.25.

- Apply Logistic regression and print the results.
- Apply Random forest and print the results.
- Apply Random Forest with hyper parameter and print the results.
- Apply XGBoost Classifier and print the results.
- Apply Decision Tree and print the result.
- Apply AdaBoost and print the result.
- Apply Ensemble classifier and print the result.
- Compare the Results.
- Exit.

IV. RESULT

Table 4.1: Comparison of Accuracy. (Test dataset)

Implemented Algorithms	**Accuracy in %**
Majority Classifier [60]	57.24
AdaBoost Classifier [60]	68.04
Logistic Regression	62.04
Random Forest	61.21
Random Forest with hyperparameters	67.77
XGBoost Model	68.47
Ensemble Classifier [60]	67.50
Decision Tree	66.51

From all the model evaluation metrics that we used on our models, we are consistently seeing that XGBoost model gives least AIC value, Highest AUC area, best training test accuracy curve.

V. CONCLUSION

By utilizing robust datasets and employing diverse machine learning techniques such as logistic regression, decision trees, random forest, XGBoost and ensemble classifiers, researchers and healthcare professionals can develop accurate and reliable predictive models for early detection and intervention.

Despite these challenges, the potential benefits of machine learning in Type-2 diabetic detection are substantial, offering the opportunity to improve patient outcomes, reduce healthcare costs, and ultimately contribute to the advancement of personalized medicine. Continued research, collaboration, and innovation in this area are essential to realizing the full potential of machine learning in diabetes care.

REFERENCES

[1] P. Dua, F. J. Doyle, and E. N. Pistikopoulos, "Model-based blood glucose control for type 1 diabetes via parametric programming," IEEE Trans. Biomed. Eng., vol. 53, no. 8, pp. 1478_1491, Aug. 2006.

[2] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes_2020," Diabetes Care, vol. 43, no. 1, pp. S14_S31, Jan. 2020.

[3] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," IEEE Trans. Biomed. Eng., vol. 54, no. 5, pp. 931_937, May 2007.

[4] S. Guerra, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, "Enhancing the accuracy of subcutaneous glucose sensors: A real-time deconvolution-based approach," IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1658_1669, Jun. 2012.

[5] J. M. Norris, R. K. Johnson, and L. C. Stene, "Type 1 diabetes_Early life origins and changing epidemiology," Lancet Diabetes Endocrinol., vol. 8, no. 3, pp. 226_238, Mar. 2020.

[6] National Diabetes Statistics Report, 2020. Accessed: Jan. 15, 2021. [Online]. Available: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[7] ID Federation. IDF DIABETES ATLAS 9th Edition 2019. Accessed: Jan. 15, 2021. [Online]. Available: https://diabetesatlas.org/en/.

[8] L. Olansky and L. Kennedy, "Finger-stick glucose monitoring: Issues of accuracy and specificity," Diabetes Care, vol. 33, no. 4, pp. 948_949, Apr. 2010.

[9] J. B. Buse, D. J.Wexler, A. Tsapas, P. Rossing, G. Mingrone,C. Mathieu, D. A. D'Alessio, and M. J. Davies, "2019 update to: Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American diabetes association (ADA) and the European association

for the study of diabetes (EASD)," Diabetologia, vol. 63, no. 2, pp. 221_228, Feb. 2020.

[10] M. Langendam, Y. M. Luijf, L. Hooft, J. H. D. Vries, A. H. Mudde, and R. J. Scholten, "Continuous glucose monitoring systems for type 1 diabetes mellitus," Cochrane Database Syst. Rev., vol. 2012, no. 1, pp. 1_144, 2012, Art. No. CD008101.

[11] C. Choleau, J. C. Klein, G. Reach, B. Aussedat, V. Demaria-Pesce, G. S.Wilson, R. Gifford, and W. K.Ward, "Calibration of a subcutaneous amperometric glucose sensor: Part 1. Effect of measurement uncertainties on the determination of sensor sensitivity and background current," Biosensors Bioelectronics, vol. 17, no. 8, pp. 641_646, Aug. 2002.

[12] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Comput. Sci., vol. 132, pp. 1578_1585, Jun. 2018.

[13] H. Kaur and V. Kumar, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Inform. vol. 16, pp. 1_11, Jul. 2020.

[14] K. Kincade, "Data mining: Digging for healthcare gold," Insurance Technol., vol. 23, no. 2, no. 2, pp. 2_7, 1998.

[15] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnology. J., vol. 15, pp. 104_116, Jan. 2017.

[16] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241_266, Oct. 2013.

[17] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, "Data-mining technologies for diabetes: A systematic review," J. Diabetes Sci. Technol., vol. 5, no. 6, pp. 1549_1556, Nov. 2011.

[18] M. Durairaj and K. Priya, "Breast cancer prediction using soft computing techniques a survey," Int. J. Comput. Sci. Eng., vol. 6, no. 8, pp. 135_145, Aug. 2018.

[19] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," Procedia Comput. Sci., vol. 47, pp. 45_51, May 2015.

[20] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," Energy Buildings, vol. 49, pp. 560_567, Jun. 2012.

[21] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, "Real-data comparison of data mining methods in prediction of diabetes in Iran," Healthcare Informat. Res., vol. 19, no. 3, no. 3, pp. 177_185, 2013.

[22] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, "Application of classification based data mining technique in diabetes care," J. Appl. *Sci.*, vol. 13, no. 3, pp. 416_422, Jan. 2013.