Early Prediction of Diabetes using Several Machine Learning Algorithms

Dr. J. A. M. Rexie
Dept. of Computer Science and Engineering
(Assistant Professor)
Karunya Institute of Technology and Sciences
(Deemed to be University)
Coimbatore, India
rexiejoseph@gmail.com

P. Santhosh
Dept. of Computer Science and Engineering
(Assistant Professor)
Karunya Institute of Technology and Sciences
(Deemed to be University)
Coimbatore, India
santhosh19@karunya.edu.in

P. Nikhil Solomon
Dept. of Computer Science and Engineering
(Assistant Professor)
Karunya Institute of Technology and Sciences
(Deemed to be University)
Coimbatore, India
nikhilsolomon0@gmail.com

P. Atul Vishnu
Dept. of Computer Science and Engineering
(Assistant Professor)
Karunya Institute of Technology and Sciences
(Deemed to be University)
Coimbatore, India
atulvishnu@gmail.com

2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) | 979-8-3503-9725-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICCS56967.2023.10142749

***Abstract*— One of the main issues diabetes poses to the medical profession globally is that its consequences are escalating swiftly. Elevated blood glucose levels cause diabetes, also referred to as diabetes mellitus or simply diabetes. On the basis of physical and chemical exams, a number of standard approaches can be used to diagnose diabetes. But, Doctors face a difficult task in predicting diabetes that affects the kidneys, eyes, heart, nerves, feet and other parts of the body.Early diagnosis analyses illness prognosis and diagnosis using a doctor's training and expertise, although this might be vulnerable to error. Machine learning and data science approaches have the potential to enrich other scientific disciplines by providing new perspectives on well-known issues. One such initiative is to assist in making predictions based on medical data. 'Machine learning' describes the processes by which computers tend to learn from experience, and is a recent area of data science. The aim of this research is to develop a system to accurately diagnose diabetes in patients at an early stage by comparing the results of several machine learning algorithms. Use four different supervised machine learning techniques: Random Forests RF, Logistic LR Regression,Decision DT Trees and (SVM) Support Vector Machines. The main objective of this research project is to develop a prognostic tool for early detection and prediction of diabetes.The model is also deployed into a web application using Python Flask, and the web application is built using HTML and CSS. Anyone can input features into the web application, and the model—which was previously developed using machine learning techniques—will then predict whether or not they will be diagnosed with diabetes.**

Keywords*—*diabetes, diagnose, prognosis

I. INTRODUCTION:

Diabetes is a rapidly spreading disease especially in children. To grasp diabetes and how it occurs, first understand what occurs to the body when it does not have diabetes. Sugar is made from the meals we eat, particularly those high in carbs. (glucose).Carbohydrates are essential for everyone, including diabetics, as they are the main source of energy for the body. Carbohydrate-rich foods include vegetables , rice , breads, dairy products, cereals, fruits, and pasta (especially starchy vegetables).When people consume these nutrients, our body converts them into glucose. Glucose is carried throughout the body through the blood.

The human brain receives some glucose in order to maintain normal functioning and thinking. Human blood cells use glucose as fuel since it remains in the blood stream. Beta cells within the pancreas create insulin. Insulin functions like a key which unlocks the door. Insulin binds to the entrance of the cell, allowing glucose to enter from the blood. Diabetes occurs The blood glucose level will rise when the pancreas produces insufficient insulin or if the body develops an immune response to insulin (insulin resistance). Diabetes causes Genetics is the primary contributing element to causes of diabetes mainly occurs due to genetic reasons. It is caused by a defect in at least two genes on chromosome 6, which control the body's response to various antigens. Viral infections affect the development of type 1 and type 2 diabetes. Studies have shown that viruses such as hepatitis B, cytomegalovirus increase the risk of diabetes, CMV, rubella, and mumps.

II. LITERATURE SURVEY:

[1].This dataset represents clinical trial data from a hospital in Luzhou, China. [1] There are 14 gunas. It randomly selects data from 68994 healthy and diabetic patients as training set. Due to data imbalance, [1] data were sampled approximately 5 times. Average results of these five tests. Minimum redundancy maximum correlation (MRMR) and Principal component analysis (PCA) were used in this study to reduce dimensionality. When all attributes are considered, the predictions of Random Forest have the best accuracy (ACC = 0.8084).

[2] The study focused on selecting relevant features and accurate classifiers to improve the accuracy of the predictive analytics. The decision tree and random forest algorithms showed high specificity, while the Naive Bayes algorithm had the highest accuracy. These findings demonstrate the effectiveness of machine learning in healthcare and can inform the development of future predictive algorithms for diabetes and other health conditions.

[3] Certain paper aim is to combine the results of various machine learning techniques to develop a system that can predict diabetes early in patients with high accuracy. Algorithms such as random forests, K nearest neighbours, support vector machines, logistic regression, and decision trees can be used. Calculate the accuracy of the model using each algorithm. The most accurate solution is selected as the solution for diabetes prediction. To achieve this goal, researcher will use various machine learning techniques to more accurately predict patient psychosis or patient diabetes. In this article, designers will apply machine learning classification and clustering techniques to a dataset to predict diabetes. Algorithms Used are Logistic Regression (LR), Gradient Boosting (GB) ,K Nearest Neighbours (KNN), Decision Trees (DT), SVM and Random Forest. The accuracy of each model varies compared to other methods.[4]. This project attempts to provide an accurate or highly accurate model that demonstrates the model's ability to accurately predict diabetes[4]. This results show that random forests outperform other machine learning algorithms in terms of accuracy. They propose a diabetes prognostic system for advanced diabetes classification that includes some extrinsic factors that may lead to diabetes in addition to conventional blood sugar, BMI, ageing, and insulin are some of the components. The new dataset boosts classification accuracy over the prior dataset[5]. In addition, pipeline method for diabetes prediction is also adopted to improve the classification accuracy[5].[6]In this study, a machine learning-based classification model was developed to diagnose diabetic patients using clinical data. Various machine learning methods were tested, including decision trees, naive Bayes, logistic regression, k-nearest neighbours, gradient boosting, random forests, and support vector machines. The model's performance was evaluated on multiple datasets, and the results showed that the suggested model could deliver improved accuracy compared to other existing studies. Depending on the dataset and the machine learning approach employed, the suggested model can achieve an accuracy improvement ranging from 2.71% to 13.13%. This research can have significant implications for the development of more accurate and efficient methods for diagnosing diabetes.

A diabetes classification system based on machine learning algorithms is described[7]. He primarily used Support Vector Machines with multiple kernels and the Diabetic dataset from the UCI Machine Repository. He found that SVMs with linear models outperform Decision DT Trees, Naive (NB) Bayes, and Neural Networks[7]. Despite this, no sophisticated comparison has been made, nor has variable selection been explored.

Classification and prediction of Diabetes is being treated with machine learning approaches[8]. To classify diabetes, the researchers used four different machine learning techniques, including Naive Bayes, AB AdaBoost, Decision DT Trees, and Random Forests.In addition to 20 experiments, they used three alternative partitioning techniques to achieve better results[8]. The National Health and Nutrition NHANES Examination Survey of diabetics and non-diabetics was used to estimate the prevalence of diabetes they made encouraging findings with the proposed methodology.On the PIMA dataset, researcher performed a comparative test of several machine learning algorithms for diabetes classification, including DT, and NB[9]. MLP outperformed other classifiers, according to the study. It has been suggested that effective feature engineering and fine-tuning can result in higher performance for MLPs[9].A diagnostic performance of 77.5% was achieved by MLP on the PIMA dataset, however, state-of-the-art comparisons are still lacking. Several health disorders, including cancer and heart disease, have been characterized using MLPs[10].

III. PROPOSED APPROACH AND ALGORITHM

A.Data preprocessing

In the machine learning data pre-processing step, one of the most important operations is dividing the dataset into training and test sets to improve the performance of machine learning models. However, understanding sampling bias can be difficult for the model, and researchers continuously strive to develop a model that performs well on both training and testing datasets.

In addition to splitting the data, efficient handling of null values and outliers is also crucial for accurate model performance. One way to handle null values is to impute them with the mean or median values of the feature. For outliers, the z-score method can be used to detect and remove them. Another approach is to use clustering techniques to identify and isolate outliers for further analysis.Moreover, scaling and normalization can also be used to ensure that all features are in the same range and have equal weight in the model. Standardization and min-max scaling are common methods used for scaling and normalization .Overall, efficient handling of null values and outliers, as well as appropriate scaling and normalization, can greatly improve the accuracy and robustness of machine learning models.

B.Why Machine Learning

Machine learning is a promising approach to predicting diabetes, given its ability to learn from labelled data and make accurate predictions on new, unseen data. Specifically, supervised machine learning is being used in this research, which involves providing input and output data to the machine learning model to learn how to correctly predict the outcome of a given situation. With labelled data, machines can learn from past examples and find a mapping function that relates input variables to output variables. Therefore, this approach is well-suited for predicting diabetes, where past patient data can be used to train a machine learning model to predict whether new patients have diabetes. In this research, four supervised machine learning models-random forests, decision trees, support vector machines, and logistic regression are being utilized to predict diabetes with high accuracy.

C.Decision trees

Decision trees are commonly used for diabetes prediction due to their ability to handle both categorical and numerical variables, as well as their interpretability. In the case of diabetes prediction, the decision tree algorithm can help identify important predictors or risk factors for diabetes, such as age, BMI, family history, and glucose levels. These important predictors can then be used to make

predictions about the likelihood of an individual developing diabetes. Furthermore, decision trees are easy to understand and visualize, making them a useful tool for healthcare professionals to communicate important information to patients.

D.Logistic Regression

Logistic regression is a well-known supervised learning machine learning technique. A categorical dependent variable is predicted by using a set of independent variables. Use logistic regression to predict outcomes for a categorical dependent variable. Hence, a separate or classified publication is required. It returns probability values between 0 and 1 rather than absolute values of 0 and 1. A valid alternative consists of either yes or no, 0 or 1, true or false, etc.

Despite similarities to linear regression and logistic regression, they are applied very differently. Classification problems and regression problems are solved using regression techniques such as logistic regression and linear regression.

This fit a "sigmoid" shaped logistic function rather than a regression line, which roughly approximates the sharp peak (0 or 1) in logistic regression. The logistic LR regression model has become a popular machine ML learning approach since it is capable of assigning probabilities and categorizing both continuous and discrete datasets. Logistic regression provides an easy way to identify which categorical variables are most important if one wants to classify observations from a variety of different types of data sets.

E.Random Forest

Good artificial intelligence algorithm Random Forest is a mechanism for tracking learning. In addition to handling classification and regression ML issues, it also supports generative models. In order to achieve this, the software uses a combination of classifiers and ensemble learning, a technique for solving problems like this one. A better understanding of how models work as well as solutions to challenging problems have been made. For improved predictive accuracy, Random Forest combines a variety of subsets of input data into an average based on a collection of decision trees. Instead of relying entirely, The Random Forest is a decision tree-based procedure which collects predictions, made by all of the decision trees, for a particular variable, and predicts the outcome of that variable based on the majority prediction. As a result, it is imperative to increase the number of trees in the forest in order to prevent overfitting and high accuracy from occurring. Since random forests combine a large number of trees to predict the classes of a dataset,Certain decision trees are capable of predicting the correct outcome, but are not able to predict it for others. Nonetheless, if all the trees are connected, then they can predict the right outcome. Therefore, the following two assumptions about the best random forest classifiers are made. A dataset's feature variables must have some actual values in the dataset to predict actual results rather than hypothetical results. The estimates of each tree should have at least one correlation

F.Support Vector Machine

For classification and regression, many supervised learning techniques are available. Machine learning classification problems use it most often, however. SVM methods allow for quick classification of new data points in the future by finding the best line or decision boundary to characterize the n-dimensional space. The correct result is the boundary hyperplane. SVM selects exponential vectors and points that contribute to the construction of the hyperplane. Because of these extreme examples of support vectors, this method is called a support vector machine.

IV. MODULE DESCRIPTION:

To run the machine learning model, this must import single web application modules such as flask and pickle, as well as specialised machine learning modules containing the functions to execute the models successfully. The following modules were utilised in this project:

A.Pandas

Described as a Python package, Pandas provides fast, flexible, and transparent data structures for handling "relational" and "labelled" data that allow you to manipulate the data in a logical and intuitive manner. It is intended as a basic, high-level building block for analysing real-world data with Python. In order to achieve its ultimate goal, its goal is to be the most powerful and flexible open-source tool for data analysis and manipulation available for any language that is open source. In that respect, it has come a long way.

B.Numpy

The NumPy Python module is a basic Python package for scientific computing. This Python library has various derived objects such as a multidimensional array object, mask arrays and matrices, and functions to perform fast array operations. Examples of these procedures include the discrete Fourier transform, basic linear algebra, basic statistical functions, stochastic simulation, and many others.

C.Sklearn

Python's Scikit-learn (sklearn) package is stable and effective. As a result of Python, a wide variety of statistical modelling and machine learning tools are readily available, including tools for classification, regression, clustering, and pre-processing of data the Data set loading and analysisis shown in “Fig 4.3.1”.

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

Fig 4.3.1

The data set has been loaded and is now ready for usage. The data set contains 8 feature values and 1 result the sample dataset is shown in “Fig 4.3.2”.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig 4.3.2

the analysis of the data set is done we can see the count mean min standard deviation etc here of the dataset.

D.Pre Processing

This section of the project is crucial because it directly affects how accurate the machine learning models will be. Here, we clean the data set by removing duplicate columns, outliers, null, and zero values so that it may be utilised for implementation.

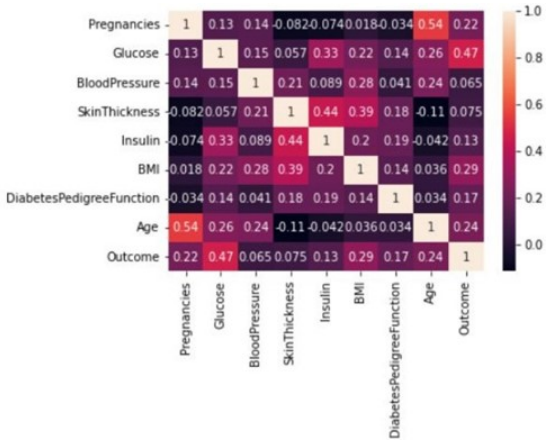

Fig 4.4.1

In “Fig 4.4.1” the heatmap of the dataset representing the correlation between each column value of the data set.

E.Machine learning model implementation

On the pre-processed dataset, we use a variety of supervised machine learning models to predict the result value and improve accuracy.

Fig 4.5.1

In “Fig 4.5.1” the above output shows the prediction of test data and its accuracy of a machine model where the accuracy score command is used to find the accuracy score of a model.

F.Taking Feature Values:

Now, using an HTML web application, we collect the feature values from the user and utilise them to feed the models we've saved with the data to make predictions about the future

Fig 4.6.1

In “fig 4.6.1” Prediction page where feature values are taken from the user.

V. CONCLUSION

The ability to predict diabetes using machine learning models has gained attention in the medical community due to the potential to identify patients at risk and provide early interventions. This study utilized patient records to develop a model that accurately predicts the presence of diabetes, while also providing insights through data analysis and visualization. The implementation of this model in a user-friendly interface allows for wider utilization by end-users with minimal coding expertise. Early prediction of diabetes is crucial due to the serious health consequences associated with the disease, including mortality. Previous literature highlights the importance of early detection and treatment of diabetes, and the proposed method offers an effective and efficient means of achieving this goal. The use of machine learning algorithms enhances the precision and accuracy of disease prediction, making it a valuable tool for clinicians and hospitals. Further research is needed to enhance the sensitivity of the model to new data, but the promising results suggest that machine learning can play an important role in diabetes prediction and prevention.

REFERENCES

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

[2] Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 6, 13 (2019).

[3] Rani, KM. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

[4] Mitushi Soni , Dr. Sunita Varma, 2020, Diabetes Prediction using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 09 (September 2020),

[5] Mujumdar, Aishwarya & Vaidehi, V.. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science.

[6] Ahmed, Nazin & Ahammed, Rayhan & Islam, Manowarul & Uddin, Md Ashraf & Akhter, Arnisha & Talukder, Md. Alamin & Paul, Bikash Kumar 2021.

[7] G. A. Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machines," 2020.

[8] Maniruzzaman, Md & Rahman, Md & Ahammed, Benojir & Abedin, Md. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems.

[9] Ahuja, Ravinder & Sharma, Subhash & Ali, Maaruf. (2019). A Diabetic Disease Prediction Model Based on Classification Algorithms.

[10] S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," 2021.

[11] C. S. Manikandababu, S. IndhuLekha, J. Jeniefer and T. A. Theodora, "Prediction of Diabetes using Machine Learning," 2022.

[12] X. Xu, X. Huang, J. Ma and X. Luo, "Prediction of Diabetes with its Symptoms Based on Machine Learning," 2021.

[13] A. C. Lyngdoh, N. A. Choudhury and S. Moulik, "Diabetes Disease Prediction Using Machine Learning Algorithms," 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Langkawi Island, Malaysia.

[14] S. M. Mahedy Hasan, M. F. Rabbi, A. I. Champa and M. A. Zaman, "An Effective Diabetes Prediction System Using Machine Learning Techniques," 2020.

[15] S. Ghane, N. Bhorade, N. Chitre, B. Poyekar, R. Mote and P. Topale, "Diabetes Prediction using Feature Extraction and Machine Learning Models," 2021.