Predicting Disease Progression in Diabetes Patients

AKSHAT GUPTA
Department Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
akshatg989@gmail.com

***Abstract* - Globally, diabetes affects 537 million adults (27-77) and this can rise up to 643 million by 2030 which makes it the deadliest non-communicable disease. There are many factors in which diabetes can affect humans such as excessive body weight, family history, physically inactive, eating excessive unhealthy food etc. Increased urination is one of the most common symptoms of diabetes. People with diabetes for a long time also have a chance of getting at risk of many diseases like kidney disease, diabetic retinopathy and nerve damage etc. But risk can be reduced if we predict diabetes at an early stage. In this research paper, a diabetes predicting system has been developed using Pima India diabetes datasets of many patients across the world and various machine learning techniques. In the study of diabetes disease is diagnosed using pre-processed data is accomplished using machine learning classification algorithm techniques. The dataset consists of a lot of patient details across the world which consist of age, gender, BMI, pregnancy, glucose, skin thickness, insulin. Our research performed this dataset will provide the highest accuracy using logistic regression and will change the face of world and reduce the number of patients infect from diabetes.**

Index Terms – Diabetes, retinopathy, logistic regression, skin thickness, diagnosed.

INTRODUCTION

Among millions of people around the world, diabetes is a major health concern. The World Health Organization (WHO) also recognizes diabetes as a major global health problem and provides guidance and information on diabetes prevention, management, and research. According to WHO's information, diabetes mellitus, commonly referred to as diabetes, affects your body's metabolism of glucose (sugar), which is a primary source of energy for your cells. In other words, Insulin resistance and Beta-cell dysfunction are two characteristics of the chronic metabolic condition known as diabetes mellitus [1]. It is described into three types: type 1 diabetes, which is an autoimmune condition that causes dysfunction in pancreatic beta cells, type 2 diabetes, which is characterized by a slow increase in insulin resistance and beta-cell dysfunction, frequently linked to obesity and certain lifestyle choices. It is becoming more common worldwide, which hurts both health and finances, and the third type is gestational diabetes linked with pregnancy because of a transient resistance to glucose uptake. The abnormality in glucose metabolism often recovers to normal after delivery. During pregnancy, Insulin secretion and release increase because insulin antagonists are created in greater numbers. A pancreas that is unable to generate enough insulin to meet this increased requirement can lead to gestational diabetes. The WHO claims that diabetes is an epidemic that is spreading across the globe and also predicts that the next decade, diabetes is expected to rank seventh in the world in terms of death rates. The World Health Organization estimated that the number of diabetic patients had increased from 108 million in 1980 to approximately 529 million in 2021. There has been a faster increase in prevalence in low- and middle-income countries than in high-income countries. Adults over the age of 40 had diabetes at a rate of 8.5% in 2014. Approximately 2 million deaths were caused directly or indirectly by diabetes in 2019. Of these deaths, 48% were among younger people. Additionally, 4,60,000 kidney disease deaths were due to diabetes and nearly 20% of cardiovascular deaths are caused by high blood glucose levels [2]. Over the past decade, diabetes fatality rates have increased by 3% as measured by age-standardized mortality rates, and in lower-middle-income countries by 13%. It is estimated that by the year 2030, the number of diabetic patients will double to 1.3 billion. The prevalence of overall diabetes in the world was 6.1% (5.8–6.5). North Africa and the Middle East now have the highest rate at the super-regional level (9.3%), and that figure is expected to rise to 16.8% by 2050[3]. Diabetes is becoming more prevalent in people's daily lives. Therefore, it is essential to research how to early and accurately diagnose diabetes mellitus early, particularly during its initial development, which is difficult for medical experts. It is possible to make preliminary diagnoses of diabetes mellitus using machine learning techniques using daily physical data, which can be used by doctors as a reference. Numerous researches have been carried out to predict diabetes automatically using machine learning techniques. Machine learning techniques enable computers to learn from experience and become more intelligent. We have used several machine learning techniques and assessed their capabilities to diagnosis diabetes. Various machine learning algorithms can be used to create a predictive model, including Naive Bayes, Decision Trees, Random Forests, Support Vector Machines, and Logistic Regression. In this research paper, datasets with various attributes have been used, including blood pressure, body mass index, age, and glucose and insulin levels. The performance of the aforementioned models has been evaluated in terms of Recall, F1-measures, Precision and lastly, based on the accuracy level. Furthermore, many researchers are getting more interested in diabetes prediction to train the program to determine if a patient has diabetes or not by applying the appropriate machine learning-based classification algorithms to the dataset. Furthermore, this paper uses machine learning techniques to predict disease progression in diabetes patients.

LITERATURE SURVEY

In recent years, predicting disease progression in diabetes patients using machine learning based classification algorithms and artificial intelligence has been gaining popularity. There is lots of work devoted to the prediction of diabetes but still it draws the attention of researchers to carry out their research work in this area of study.

Kopitar, L. Kocbek, P., Cilar, L. *et al* [5] have applied the idea of artificial intelligence to increase the precision of disease prediction. Preprocessing, feature selection, and feature categorization are the three processes that are used up in this procedure. For feature selection, techniques including k-means clustering, the genetic algorithm (GA), the harmony search algorithm (HSA), and the particle swarm optimization (PSO) are used. KNN has been used in classification techniques. Sensitivity, specificity, recall, and precision are some of the criteria used to evaluate accuracy prediction. A 91.65% accuracy rate is attained.

Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe [6] investigated the diseases of diabetes with the help of three machine learning algorithms. Predicting an individual's diabetes status can be done using SVMs (support vector machines), CNNs (convolutional neural networks), or random forests (RFs). To predict best accuracy, the experiment was repeated more than 10 times and the average accuracy was taken.

Zou Quan *et al* [7] used decision trees, random forests, and neural networks to diagnosis diabetes progression. Data was collected from Luzhou physical examinations in China. In order to reduce the dimension of the dataset, PCA was applied. And they chose various types of machine learning (ML) methods and performed independent tests on them to determine whether the method is generally applicable.

Ayon, Safial & Islam [8] have applied 5-fold and 10-fold cross-validations for training a deep learning network to diagnose diabetes. Pima Indian data was used for this analysis. A 10-fold cross-validation method was employed to find that prediction accuracy was 98.35%.

Deepti Sisodiya, Dilip Singh Sisodiya [9] have described a algorithm for predicting diabetes. The purpose of this study is to develop a model that can accurately predict diabetes. In order to predict diabetes at an early stage, they used three different classification algorithms: Decision Trees, Naive Bayes, and Support Vector Machines. Several estimation methods were used to evaluate the three models, including Accuracy, Precision, F-Measure, and Recall.

A. Nur Ghaniaviyanto Ramadhan *et al* [10] have mainly focused on data preprocessing, in order to improve the feature set, the data process must be balanced, missing values are removed, and features are enhanced. Random Forest and Logistics Regression are used for classification algorithms. Comparing pre-processed data with raw data, the outcome is 24% higher for recall and 20% higher for accuracy.

Jackins, V.Shanmuganathan et.al [11] demonstrated a multi-disease diagnosis system, including diabetes using machine learning classification algorithms. It appears that the Random Forest model is more accurate than the Naive Bayes classifier for the multi diseases.

Muhammad Azeem, Nasir kamal et.al [12] have applied six different machine learning techniques. In this study, the performance and accuracy of techniques are evaluated and compared. Based on a comparison of a variety of machine learning approaches, it can be determined which approach is most suitable for diabetes prediction.

METHODOLOGY

The methodology of diabetes prediction using machine learning involves several steps that utilize various techniques to effectively analyse and predict diabetes outcomes. This research field has gained significant attention in recent years due to the increasing prevalence of diabetes and the abundance of healthcare data available for analysis. An outline of the working processes and implementation of various machine learning algorithms are presented in this section. Fig.3.1 illustrates that how this research was conducted. Ultimately, the purpose of this paper is to investigate a model that is more accurate at predicting diabetes. In order to make predictions, we experimented with different machine learning classification algorithms.

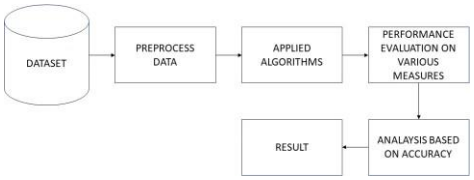

Figure 3.1 Summary of the research work

3.1 Dataset:

The data are gathered from a repository at UCI titled Pima Indian Diabetes Dataset. It contains 2000 patients with many attributes. Table 3.1 shows a dataset description of the Open Source Pima Indian Dataset.

S.no	Attributes
1.	Pregnancies
2.	Glucose
3.	Blood Pressure
4.	Skin Thickness
5.	Insulin
6.	BMI
7.	Diabetes Pedigree Function
8	Age
9.	Outcome(0/1)

Table 3.1 shows dataset description.

In the 9th attribute, there is a class variable for each data point. As indicated by this class variable, diabetics have the outcomes 0 or 1, which indicates whether they are positive or

negative. Our model is used to predict diabetes, but the dataset was a bit imbalanced: approximately 66% classes were categorized as negative which means no diabetes, while the remaining 34% were categorized as positive which means a person having diabetic Figure 3.2.

3.2 Data Preprocessing:

Among the most important procedures is data preprocessing. Usually, healthcare data is contaminated with missing values and other contaminants, reducing its effectiveness. The purpose of data preprocessing is to increase the quality and effectiveness of mined results. By using machine learning classification techniques on a dataset, this method can yield accurate results and good predictions. We must preprocess the Pima Indian diabetes dataset in two phases.

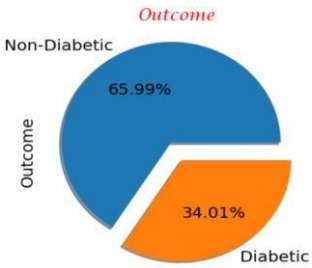

Figure 3.2 Percentage ratio of Diabetic and Non-Diabetic Patient

3.2.1 Missing Values removal:

Getting rid of missing values (also known as handling with missing data) is important step in data preprocessing. Missing values may have an adverse effect on quality and accuracy of our results. Depending on the nature of our data, there are numerous approaches to handle missing values. Removes all instances with a value of zero (0). Getting zero (0) as worth is not feasible. Therefore, this case is removed. We construct feature subsets by removing unnecessary features, a process known as features subset selection, which decreases data dimensionality and allows us to work quicker. Another approach, replace missing values in the same column with the mean, median, or mode of the non-missing values. This is a straightforward way for preserving the general distribution of the data.

3.2.2 Splitting of data:

After handling missing values in our dataset, the next step is typically to divide our data into training and testing sets. By dividing the data into training and testing sets, we can assess how well our model performs on unseen data. Splitting data is used to determine a model's performance and generalization capability. In this research, we are splitting the data into 80% for the train and 20% for the test. This splitting of data can be done in Python using libraries such as scikit-learn. The following illustrations illustrate this:

- from sklearn.model_selection import train_test_split
  X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.3, random_state=42)
- X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)

3.3 Applied Algorithms:

The prediction of diabetes is based on different classification. A main goal of this study is to apply machine learning algorithms to analyse the performance and accuracy of these methods, as well as to identify the major features responsible for the accuracy of these methods. The following algorithms used in this research. They are:

3.3.1 Logistic Regression:

Although it has several drawbacks, such as the assumption of linearity and the restricted representation of complicated connections, logistic regression is still a popular and useful technique in many applications, including the analysis of the diabetes dataset. It is a well-liked option in the context of the diabetes dataset because of its capability to handle binary classification tasks, interpretability, efficiency, and applicability as a baseline model.

Without any doubt, it is employed in tasks requiring binary categorization. It is intended especially for scenarios in which the dependent variable can be classified into two categories, such as “yes/no,” “true/false,” or, in the case of the diabetes dataset, “diabetic/non-diabetic.” The objective of logistic regression is used to calculate the likelihood that an observation belongs to a certain class based on one or more independent variables. A probability value between 0 and 1 that represents the results of logistic regression may be understood as the likelihood of falling into the positive class (for example, having diabetes). The logistic regression model converts a linear combination of the input features into the predicted probability using a logistic function, also known as a sigmoid function. The logistic function is defined as:

$$P(y = 1|X) = 1 / (1 + e^{-(b0 + b1x1 + b2x2 + ... + bnxn)})$$

Where,

$P(y = 1|X)$ is the probability of the positive class given the input features X,

e, the natural logarithm which is approximately 2.718,

b0, b1, b2,..., and bn is the logistic regression model's coefficients,

x1, >- x2,..., >- xn are the corresponding feature values.

3.3.2 Support Vector Machines:

Support Vector Machine (SVM) is a supervised machine learning technique. SVM is the most often used classifier. SVM generates a hyperplane that divides two classes. It may generate a hyperplane or series of hyperplanes in three dimensions. Although, SVMs offer benefits in dealing with nonlinear data, outliers, and high-dimensional spaces. SVMs are computationally demanding and may need careful parameter tweaking. Furthermore, SVMs may not perform optimally when the dataset is skewed or there is a high degree of overlap across classes. Because of its capacity to handle nonlinear data, resistance to outliers, effectiveness in high-dimensional spaces, generalization performance, and binary

classification capabilities. That's why, SVMs are a preferred choice for the diabetes dataset.

3.3.3 Decision Tree Classifier:

The selection of a predictive model for diabetes prediction should be based on the unique dataset, the complexity of the interactions involved, and the trade-off between interpretability and predictive accuracy. Decision trees may be an effective beginning point for developing a prediction model and can be part of a larger study that includes more sophisticated approaches when necessary. A decision tree is used when the answer variable is categorical. It is a tree-like architecture that represents the classification process based on input features.

3.3.4 K- Neighbours Classifier:

K-Nearest Neighbours (KNN) is a straightforward yet powerful supervised machine learning technique for classification and regression applications. KNN is very beneficial for prediction jobs in which we wish to generate predictions based on the similarity between new and existing data points in our dataset. K-Nearest Neighbours (KNN) can be used to predict diabetes, particularly for binary classification tasks in which we want to predict whether a person has diabetes (positive class) or does not have diabetes (negative class) based on specific traits or risk factors.

RESULT AND DISCUSSION

A correlation matrix displays the coefficients of correlation between variables. Each cell represents a relationship between two variables. In addition to describing data, it can also be incorporated into more complex analyses or used as a diagnostic tool for future studies. We can see how several properties are connected to one another in this fig 4.1.

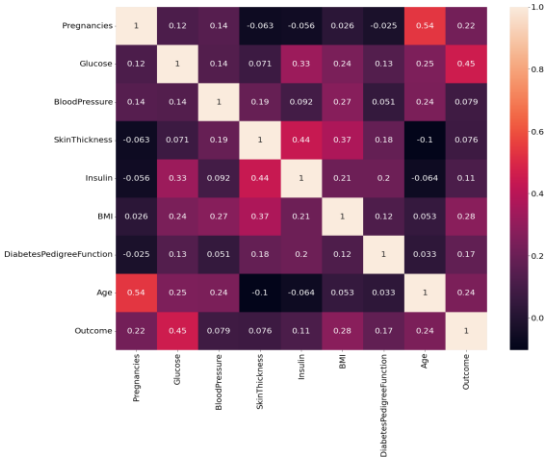

Figure 4.1. Correlation matrix for correlation analysis.

In this section, we used several performance metrics such as precision, recall, f1-score, and accuracy to evaluate the quality of a model's predictions. The equation of various performance metrices are expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True positives} + \text{False positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True positives} + \text{False negatives}}$$

$$\text{F1 - Score} = \frac{2*\text{Precision}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{Total Number of Predictions}}{\text{Number of correct Predictions}}$$

Classification Report of logistic regression:

	precision	recall	Fi-score	support
0.0	0.77	0.92	0.84	90
1.0	0.83	0.58	0.68	59
accuracy			0.79	149
Macro avg.	0.80	0.75	0.76	149
Weight avg.	0.79	0.79	0.78	149

Table 4.1 shows Logistic regression.

Classification Report of support vector machine:

	precision	recall	Fi-score	support
0.0	0.76	0.90	0.83	90
1.0	0.79	0.58	0.67	59
accuracy			0.77	149
Macro avg.	0.78	0.74	0.75	149
Weight avg.	0.77	0.77	0.76	149

Table 4.2 shows Support Vector Machine.

Classification Report of decision tree:

	precision	recall	Fi-score	support
0.0	0.69	0.77	0.73	90
1.0	0.57	0.47	0.52	59
accuracy			0.65	149
Macro avg.	0.63	0.62	0.62	149
Weight avg.	0.64	0.65	0.64	149

Table 4.3 shows Decision Tree.

Classification Report of K- nearest neighbor:

	precision	recall	Fi-score	support
0.0	0.73	0.82	0.77	90
1.0	0.66	0.53	0.58	59
accuracy			0.70	149
Macro avg.	0.69	0.67	0.68	149
Weight avg.	0.70	0.70	0.70	149

Table 4.4 shows K-Nearest Neighbor.

This research went through several stages. We propose a technique that employs multiple classification algorithms. These are common Machine Learning techniques for getting precise accuracy from data. In this study, the logistic

regression classifier beats the other classifiers. It gets a reasonable accuracy of 78.52% on the test data, as well as decent precision, recall, and F1-scores in both classes. However, there is definitely space for improvement, notably in the diabetic class's recall and F1-score (1.0). To improve the model's prediction capabilities on the diabetes dataset, more analysis and model tuning may be required. Overall, we applied the best Machine Learning approaches for prediction and best performance accuracy. The figure 4.2 depicts the outcome of several Machine Learning approaches.

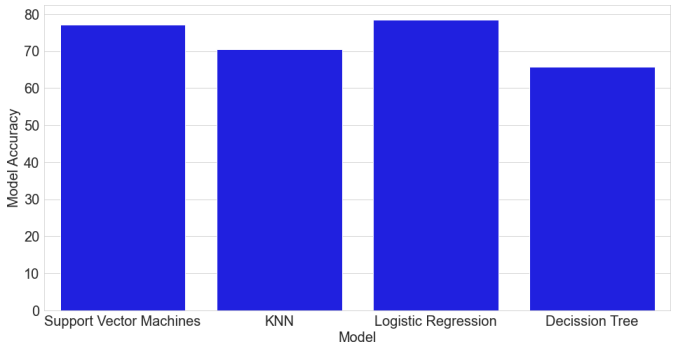

Figure 4.2. Accuracy Result of Machine learning methods

CONCLUSION

Diabetes is one of the main reasons for reducing life expectancy and quality and predicting this dangerous disease at an early stage increases the quality of life and also reduces the risk of many diseases in the long run. In this paper, Predicting Disease Progression in Diabetes Patients various machine learning approaches have been proposed. Logistic regression, support vector machine, decision tree classifier, and k-neighbor preprocessing technique are used to measure accuracy and correct precision. The research paper reported various performance metrics such as recall, F1 score, support, and ensemble technique. In this research paper, we achieve the highest accuracy with a logistic regression approach. Next this technique has been applied to demonstrate the varsity of the proposed prediction system. Finally, this technique can be launched in the form of a website only to experts of this field. There are some further scopes of this work. For example, we get additional data and launch it in the form of an application so non-expert can also analyse them in just a few minutes and we also use other machine learning techniques to make it more accurate.

REFERENCES

[1] Dunne, P., Dunne, P., Kwasnicka, D., Byrne, M., & McSharry, J. (2021, December 14). Barriers and enablers to sustaining self-management behaviors after completing a self-management support intervention for type 2 diabetes: a protocol for a systematic review and qualitative evidence synthesis. https://scite.ai/reports/10.12688/hrbopenres.13466.1

[2] Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019. Results. Institute for Health Metrics and Evaluation. 2020 (https://vizhub.healthdata.org/gbd-results/).

[3] Lancet 2023; 402: 203–34 Published Online June 22, 2023 https://doi.org/10.1016/ S0140-6736(23)01301-6.

[4] American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12- s064

[5] Kopitar, L., Kocbek, P., Cilar, L. *et al.* Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep* 10, 11981 (2020). https://doi.org/10.1038/s41598-020-68771-z

[6] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.

[7] Zou Quan, Qu Kaiyang, Luo Yamei, Yin Dehui, Ju Ying, Tang Hua(2018). Predicting Diabetes Mellitus With Machine Learning Techniques ,Frontiers in Genetics ,volume-9, doi: 10.3389/fgene.2018.00515.

[8] Ayon, Safial & Islam, Md. (2019). Diabetes Prediction: A Deep Learning Approach. International Journal of Information Engineering and Electronic Business. 11. 21-27. 10.5815/ijieeb.2019.02.03.

[9] Sisodia, Deepti & Sisodia, Dilip. (2023). Prediction of Diabetes using Classification Algorithms.

[10] Ramadhan, Nur Ghaniaviyanto & Adiwijaya, Kang & Romadhony, Ade. (2021). Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest. International Journal of Advanced Computer Science and Applications. 12. 2021. 10.14569/IJACSA.2021.0120726.

[11] Jackins, V. & Shanmuganathan, Vimal & Kaliappan.,M.E.,Ph.D, Dr.M & Lee, Mi. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing. 77. 1-22. 10.1007/s11227-020-03481-x.

[12] Sarwar, Muhammad Azeem & Kamal, Nasir & Hamid, Wajeeha & Shah, Munam. (2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. 1-6. 10.23919/IConAC.2018.8748992.