Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 1, August : 2023

DIABETES REPORT ANALYZER

Dnyaneshwar Sonawane Department of Computer Science & Engineering SKN Sinhgad Institute of Technology & Science, Lonavala Pune, India yaneshwarsonawane.sknsits.comp@ gmail.com
Prerna Bodakhe Department of Computer Science & Engineering SKN Sinhgad Institute of Technology & Science, Lonavala Pune, India prernabodakhe.sknsits.comp@gmail.co m
Vaishnavi Punalkar Department of Computer Science & Engineering SKN Sinhgad Institute of Technology & Science, Lonavala Pune, India vaishnavisp2323@gmail.com
Asst. Prof. Ravishankar C. Bhaganagare Department of Computer Science & Engineering SKN Sinhgad Institute of Technology & Science, Lonavala Pune, India rcbhaganagare.sknsits@sinhgad.edu.
Raj pandharpatte Department of Computer Science & Engineering SKN Sinhgad Institute of Technology & Science, Lonavala Pune, India rajpandharpatte.sknsits. comp@gmail.com

Abstract
Diabetes is a chronic disease that has the potential to become a significant global health crisis. The International Diabetes Federation estimates that there are currently 382 million people living with diabetes worldwide, and this number is expected to double by 2035 to 592 million. The disease is characterized by high levels of blood glucose, which can cause symptoms such as increased thirst, frequent urination, and increased hunger. Diabetes is also a leading cause of complications such as blindness, kidney failure, amputations, heart failure, and stroke. The human body turns food into glucose for energy, and insulin, produced by the pancreas, is necessary for cells to use glucose. In people with diabetes, this system does not work correctly. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other types, such as gestational diabetes, which occurs during pregnancy. Machine learning is a rapidly growing field in data science that deals with the ways in which machines can learn from experience. The objective of this project is to develop a system that can accurately predict the onset of diabetes in patients. To achieve this, the project combines the results of various machine learning techniques, including K nearest neighbor, logistic regression, random forest, support vector machine, and decision tree. The accuracy of the model using each of these algorithms is calculated, and the one with the highest accuracy is selected as the model for predicting diabetes.

Keywords— Machine Learning, K-nearest neighbor, Support vector machine, Accuracy, Diabetes

I. INTRODUCTION

Diabetes is a rapidly growing disease affecting people of all ages, including young adults. To understand diabetes and its development, it is essential to comprehend what occurs in a non-diabetic body. Carbohydrate foods, such as bread, pasta, rice, fruits, dairy products, and starchy vegetables, are the primary source of sugar (glucose) in the body. After consumption, the body breaks down these foods into glucose, which moves around in the bloodstream. Glucose provides energy to the brain and other cells in the body, and the excess glucose is stored in the liver for future use. Insulin, produced by beta cells in the pancreas, is required for the body to utilize glucose for energy. Insulin acts as a key to the cells, allowing glucose to enter the cells from the bloodstream. However, if the pancreas does not produce enough insulin or the body cannot use the insulin it produces, glucose builds up in the bloodstream, causing hyperglycemia and leading to diabetes mellitus.

Diabetes Mellitus is characterized by high levels of sugar (glucose) in the bloodstream and urine. There are three main types of diabetes: Type 1, Type 2, and Gestational. Type 1 diabetes occurs when the immune system fails to produce enough insulin, and there is no known prevention or cure. Type 2 diabetes results from the cells producing insufficient insulin or the body's inability to use insulin

correctly. This type of diabetes is caused by both genetic and lifestyle factors and affects 90% of individuals with diabetes. Gestational diabetes occurs in pregnant women who suddenly develop high blood sugar, and there is a high chance that Type 1 or Type 2 diabetes may develop after a pregnancy affected by Gestational diabetes.

Common symptoms of diabetes include frequent urination, increased thirst, tiredness/sleepiness, weight loss, blurred vision, mood swings, confusion, difficulty concentrating, and frequent infections. Genetic factors are the primary cause of diabetes, with at least two mutant genes on chromosome 6 affecting the body's response to various antigens.

The Diabetic Report Analyzer is a machine learning- based tool that aims to address this challenge by analyzing EHRs of patients with diabetes. The Diabetic Report Analyzer can extract relevant information from EHRs, such as medications, lab test results, and clinical notes, and use this information to provide clinicians with insights into the patient's condition and personalized recommendations for treatment.

In this report, we present the development and evaluation of the Diabetic Report Analyzer, a machine learning-based tool for analyzing EHRs of patients with diabetes. We describe the data sources and methods used to develop the Diabetic Report Analyzer, as well as the results of our evaluation of the tool's performance. We also discuss the potential impact of the Diabetic Report Analyzer on diabetes management and the challenges associated with implementing machine learning-based tools in healthcare. Overall, this report aims to contribute to the growing body of literature on the use of machine learning in healthcare and its potential to improve patient outcomes.

II. LITERATURE SURVEY

First, Yasodha et al. [1] have proposed a classification system for different types of datasets to determine whether a person has diabetes or not. The dataset used in this study consists of 200 instances with nine attributes, collected from a hospital warehouse, which are divided into two groups based on blood tests and urine tests. The WEKA tool is used to classify the data, and 10-fold cross-validation is used to evaluate the performance of the classifiers on this small dataset. Naïve Bayes, J48, REP Tree, and Random Tree classifiers are used in this study, and their results are compared. The study concludes that J48 performs the best, with an accuracy of 60.2%, compared to the other classifiers.

The study by Aiswarya et al. [2] aims to develop a quicker and more efficient method of identifying diabetes by analyzing patterns in the data using classification analysis through Decision Tree and Naïve Bayes algorithms. By using the PIMA dataset and cross-validation approach, the study found that J48 algorithm has an accuracy rate of 74.8%, while the Naïve Bayes algorithm has an accuracy of 79.5% using 70:30 split.

Gupta et al. [3] aimed to compare the accuracy, sensitivity, and specificity of several classification methods in WEKA, Rapidminer, and Matlab using the same parameters. They applied JRIP, Jgraft, and BayesNet algorithms and found that Jgraft had the highest accuracy of 81.3%, sensitivity of 59.7%, and specificity of 81.4%. The study also concluded that WEKA performed better than Matlab and Rapidminer.

Lee et al. [4] focused on applying the decision tree algorithm called CART on the diabetes dataset after applying the resample filter to handle the class imbalance problem. The author emphasized the need to handle the class imbalance problem to achieve better accuracy rates and to boost the accuracy of the predictive model during data preprocessing stage.

III. METHODOLOGY

In this section, we will discuss the different classifiers used in machine learning for predicting diabetes, and we will also introduce our proposed methodology for improving accuracy. We employed five distinct methods in this study, and their definitions are explained below. The accuracy metrics of the machine learning models are the output, which can then be used for prediction.

A. Dataset Description

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 1, August : 2023

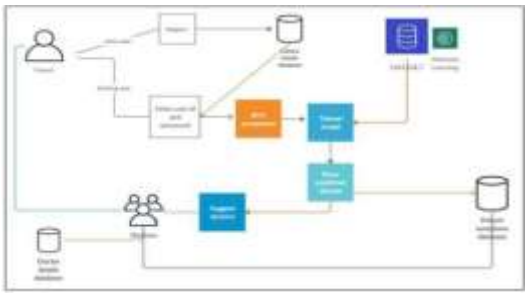
The diabetes dataset used in this study consists of 2000 cases and was obtained from https://www.kaggle.com/johndasilva/diabetes. The objective of the dataset is to predict, based on various measures, whether a patient has diabetes or not.

- The dataset for diabetes contains 2000 data points, each with 9 features.
- The feature we are interested in predicting is called "Outcome", where a value of 0 represents no diabetes and a value of 1 represents the presence of diabetes.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

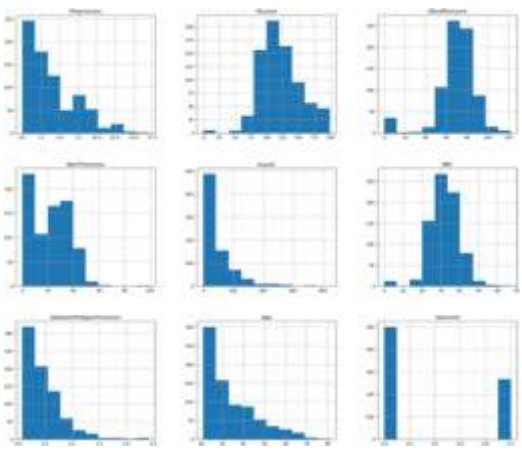
- Null Values are not present in dataset.



Proposed Architecture Diagram

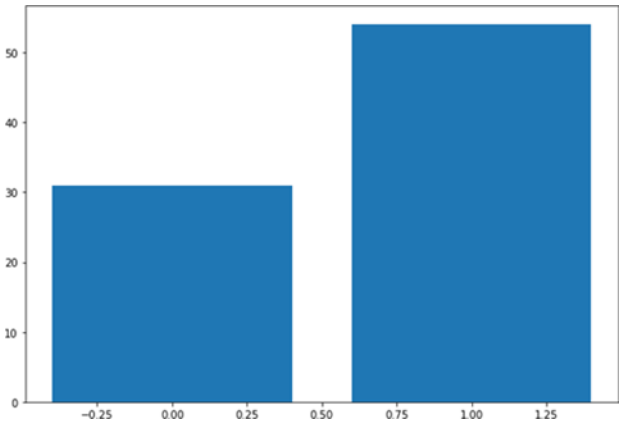
IV. RESULT AND DISCUSSION

Histogram:



To gain a better understanding of the data distribution, we can examine the histograms. The histograms illustrate how each feature and label are distributed over different ranges, further underscoring the necessity for scaling. Additionally, whenever discrete bars are present, they indicate that the variable is categorical, requiring preprocessing before applying machine learning techniques. The outcome labels comprise two classes, namely 0 indicating the absence of disease and 1 indicating the presence of disease. It is crucial to handle these categorical variables and preprocess the data to ensure accurate and meaningful predictions.

Bar Plot For Outcome Class



Outcome Class

The graph above indicates that the dataset is skewed towards data points with an outcome value of 0, which indicates the absence of diabetes. The number of non- diabetic patients is nearly double the number of diabetic patients.

K- Nearest Neighbors:

K-Nearest Neighbors (KNN) is a type of supervised learning algorithm used for classification and regression problems. In KNN, the K refers to the number of nearest neighbors that a data point is compared to in order to determine its classification. The algorithm works by calculating the distance between a data point and all other

data points in the training set. Then, the K nearest data points (nearest neighbors) are selected based on their calculated distances to the query data point. Finally, the classification or regression of the query data point is determined based on the majority vote or average value of its K nearest neighbours. One of the advantages of KNN is that it is a non-parametric algorithm, meaning it does not make any assumptions about the distribution of the data. Additionally, KNN is a simple and easy to understand algorithm. However, one of the main disadvantages of KNN is that it can be computationally expensive and slow for large datasets, especially when calculating the distance between all data points.

Moving on to the k-Nearest Neighbor (k-NN) algorithm, it is considered one of the simplest machine learning algorithms. The model building process involves only storing the training dataset. When a new data point needs to be predicted, the algorithm searches for the nearest neighbors in the training dataset. These neighbors are the data points that are closest to the new data point.

Training Precision	0.82
Testing Precision	0.79

Support Vector Machine:

The SVM classifier is a popular algorithm that is used for classification tasks. It works by creating a hyperplane that maximally separates the classes by adjusting the distance between the data points and the hyperplane. The distance between the data points and the hyperplane is determined using kernel

functions, which are a set of mathematical functions used to transform the data into a higher-dimensional space. There are several types of kernels, such as linear, polynomial, and radial basis function (RBF), which are used to decide the hyperplane. The SVM classifier has been widely used in various applications, such as image classification, text classification, and medical diagnosis.
Support Vector Machine (SVM) is a popular supervised machine learning algorithm used for classification and regression analysis. SVM tries to find the best possible boundary between the data points of different classes by creating a hyperplane in a high-dimensional space. The objective is to maximize the margin, which is the distance between the hyperplane and the closest data points from each class.
The SVM algorithm can handle linear and non-linear classification problems by using different kernel functions such as linear, polynomial, and radial basis function (RBF). The kernel function transforms the data into a higher- dimensional space, where it is easier to separate the classes. SVM is known for its ability to handle complex datasets with a large number of features. It is widely used in various applications such as image classification, text classification, and bioinformatics. SVM has also been used in medical applications, including the prediction of diabetes and cancer diagnosis.

V. CONCLUSION AND FUTURE WORK

The primary goal of this study is to develop and implement a machine learning-based method for diabetes prediction and to evaluate its performance. The proposed method involves the use of SVM, KNN, and logistic regression algorithms. This approach could potentially provide clinicians with a useful tool for making better decisions regarding disease status. Early detection of diabetes is a crucial medical issue, and this study aims to systematically design a system that can accurately predict the disease. The study evaluates the performance of five different machine learning classification algorithms using various measures, with experiments conducted on the John Diabetes Database. In the future, the designed system and classification algorithms could potentially be applied to other diseases. The study also suggests that the system could be improved and extended by automating diabetes analysis with other machine learning algorithms. The researchers hope that the system will achieve an accuracy rate of over 98%.

ACKNOWLEDGMENT

I am grateful to Dr .M.S.Rohokale and Prof .Ravishankar
C. Bhaganagare (Assistant Professor), Department of Computer Science & Engineering at SKN Sinhgad Institute of Technology & Science, Lonavala, for their valuable guidance, help, cooperation, and encouragement.
I would like to extend my gratitude to SKN Sinhgad Institute of Technology & Science, Lonavala College for providing me with this opportunity to enhance my knowledge and skills in Machine Learning. I am also thankful to my parents and family members for their unwavering support, both morally and economically.
This acknowledgement would be incomplete without expressing my heartfelt thanks to everyone who has contributed directly or indirectly to this work. Any inadvertent omission is purely unintentional and does not reflect a lack of gratitude on my part.

REFERENCES

[1] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.
[2] Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[3] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-
11933-5.
[4] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.
[5] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.
[6] Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.
[7] Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 , 554–559doi:10.1109/CICN.2010.109.
[8] Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–
1038.
[9] https://www.kaggle.com/johndasilva/diabetes.
[10] Dr. M. Renuka Devi and J. Maria Shyla, “Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1
(2016).
[11] VeenaVijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, “A Machine Learning Approach” ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)| 10- 12 December
2015 | Trivandrum.
[12] Santhanam, T. and Padmavathi, M.S., 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for a diabetes diagnosis .Procedia Computer Science, 47, pp.76- 83.(2015).
[13] S. Habibi, M. Ahmadi, S. Alizadeh Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining Glob J Health Sci, 7 (5) (Mar 18 2015), pp. 304-310, 10.5539/gjhs.v7n5p304.
[14] 44A. Allalou, A. Nalla, K.J. Prentice, Y. Liu, M. Zhang, F.F. Dai, et al. A predictive metabolic signature for the transition from gestational diabetes to type 2 diabetes Diabetes (Jun 23 2016) [pii: db151720. [Epub ahead of print]]