IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 8, April 2023

Diabetes Prediction using Machine Learning

Ms. P. V. Deshmukh[1], Ashwini Ghate[2], Prajakta Mathe[3], Aditi Dhote[4], Pratiksha Patte[5], Vrushali Mange[6]

Assistant Professor, Department of Computer Science and Engineering [1]
Under Graduate Students, Department of Computer Science and Engineering[2,3,4,5,6]
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

Abstract: *High levels of glucose in the bloodstream lead to the development of diabetes , which results in frequent urination, increased thirst, and increased hunger. It is crucial to address diabetes promptly as untreated cases may lead to severe complications in various body organs such as the heart, kidneys, blood pressure, and eyes. Predictive analytics over big data is a challenging task, particularly in healthcare. However, it can aid healthcare practitioners in making quick decisions about patients' health and treatment based on big data. The performance and accuracy of ML algorithms used in predictive Data analysis for predicting the occurrence of diabetes are compared and analyzed across various disciplines. In this study, different classification Computational methods, which may involve various algorithms, such as SVM, KNN, Logistic regression, and Random forest, were considered, and their performance metrics such as Recall, F-Measure, Precision, and Accuracy were evaluated Derived from the confusion matrix. According to the experimental results, the SVM and ontology classifiers yielded the highest accuracy for diabetes prediction.*

Keywords: ML, Diabetes Prediction, SVM, KNN, Logistic Regression (LR), Random Forest.

I. INTRODUCTION

Diabetes prevalent condition in today's world and poses significant Challenges are present in both developed & developing nations. When we eat, the insulin hormone The pancreas produces a substance that permits glucose to enter the bloodstream. Pancreatic dysfunction leads to diabetes can result in several serious conditions such as coma, retinal failure , renal, destruction of pancreatic beta cells, dysfunction of the cardiovascular and cerebral vascular systems, peripheral vascular diseases, sexual and joint dysfunction, weight loss, ulcers, and negative effects on immunity. Diabetes ranks as the third leading cause of death, trailing behind heart disease and cancer. However, with the advancement of machine learning technologies, we may be able to tackle this problem. Machine learning and data mining aim to extract information from data and produce clear and understandable representations. Our goal is to utilize ML to develop the diabetes diagnosis system capable of predicting whether a Whether or not the patient has diabetes. Obesity, high blood glucose levels, and other factors can contribute to diabetes, which affects insulin and carbohydrate metabolism, leading to abnormal levels of glucose in the blood. Insufficient production of insulin by the body leads to the development of diabetes . An organization known as the World Health Organization(WHO) estimates that approximately 422 million individuals globally worldwide with the majority of individuals with diabetes reside in low or Nations with incomes considered to be in the middle range. Diabetes is categorized into type one and type two. Type one characterized by a lack of insulin production, while type two diabetes is characterized by inadequate insulin response and production.

II. LITERATURE SURVEY

Arwatki Chen Lyngdoh et al. conducted research on predicting diabetes disease using 5 supervised ML Algo: KNN, Naive Bayes, Decision Tree Classifier, Random Forest, and SVM. by including current risk variables and performing cross-validation, they achieved consistent accuracy with the KNN classifier achieving a high accuracy of 76%. The main objective of the study was to identify the best outcomes for accurately predicting diabetes disease, considering accuracy and computing time.

Mitushi Soni et al. ML classification and ensemble techniques were employed to make predictions about diabetes using a dataset. They employed K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest algorithms, and found Random Forest outperformed the others in terms of accuracy.

DOI: 10.48175/IJARSCT-9556

Sivaranjani S et al. used SVM and Random Forest(RF)methods for identifying potential risks of Diabetes Related Diseases. After data preprocessing and implementing forward & backward stepwise feature selection was utilized to identify the most impactful features, they employed. Principle Component Analysis was employed to reduce dimensionality. Their study , which outperformed Support Vector Machine's 81.4% accuracy.

Shejal Kale et al. applied MLClassification & Using ensemble techniques to make predictions about diabetes on a given dataset.They utilized KNN, Logistic Regression(LR), Decision Tree(DT) , SVM , Gradient Boosting(GB) , and Random Forest(RF) algorithms, and found that (RF) Random Forest had the best accuracy.

Ashwini R et al. trained ML ALGO such as KNN, Random Forest(RF), Logistic Regression(LR), and SVM using various datasets. They used preprocessing techniques to improve the accuracy of their models and prioritized risk factors by employing various feature selection approaches.

III. METHODOLOGY

The aim of the project is to enhance the accuracy of diabetes prediction models. Our approach involved exploring various ML ALGOS like KNN , for classification and prediction. In the subsequent sections, we provide a concise overview of our methodology.

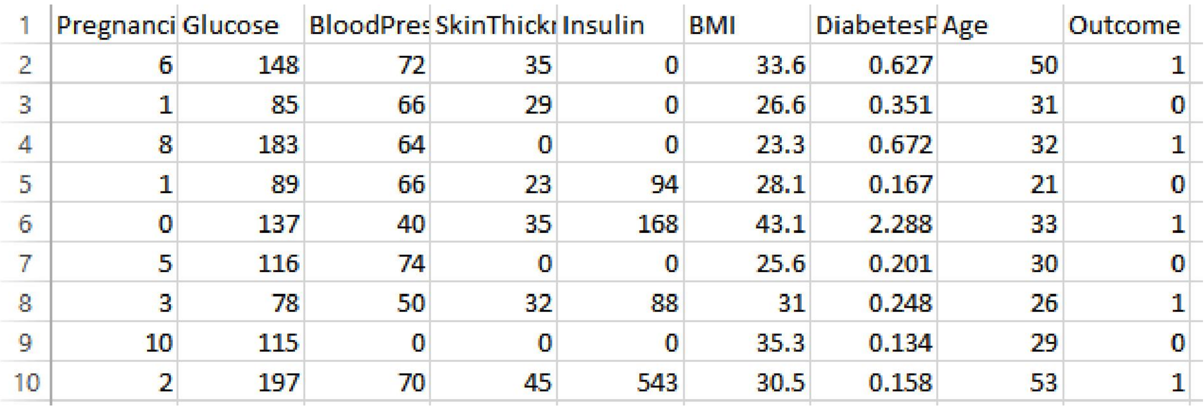
Description of the Dataset

The data utilized in this project was got From the UCI Machine Learning repo. and is known as Pima Diabetes CSV File. It comprises several features of 768 patients.

The ninth attribute in each data point represents the class variable, which indicates whether the individual is -or+for diabetes, denoted by 1 and 0, respectively.

Sr. Number	Attributes
I	Pregnancy Attribute
II	Glucose Attribute
III	Blood Pressure Attribute
IV	Skin thickness Attribute
V	Insulin Attribute
VI	Body Mask Index
VII	Diabetes Pedigree Function Attribute
VIII	Age Criteria

Table 1: Dataset Contents



1	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Figure 1 : Pima Indians Dataset

DOI: 10.48175/IJARSCT-9556

Distribution of Diabetic Patient

In our attempt to develop a diabetes prediction model, we encountered a slightly imbalanced dataset. Out of the total 768 samples, around 500 were Designated as 0, denoting the nonexistence of diabetes., while 268 were designated as 1, denoting the existence of diabetes.

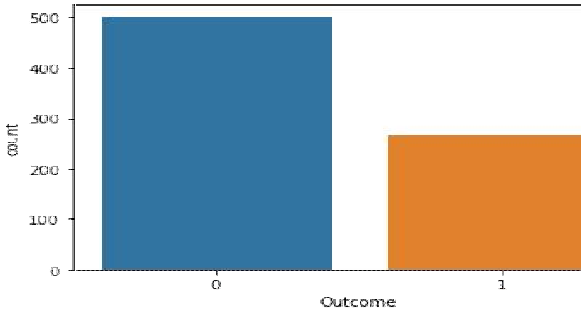


Figure 2: The proportion of patients with diabetes compared to those without diabetes.

1. **Data Pre-processing** - The process of preprocessing data is of utmost importance, especially for data concern with healthcare, which may contain missing values and Other contaminants that may affect the effectiveness of data mining. This process is essential to achieve accurate results and successful predictions with the help ofML methodology on the CSV file. To work with the Pima Indian diabetes dataset, we require, preprocessing in couple of steps.
2. **Missing Values Elimination**- A value of zero are removed since it is not possible to have a value of zero for certain features. This process helps in feature subset selection by eliminating irrelevant features or instances, reducing the dimensionality of the data and enabling faster processing.
3. **Categorization of data**- The data is in normal form and divided into training and testing sets after undergoing cleaning. The algorithm is trained on the training dataset, and the test dataset is kept aside. This training process produces a model based on logic, algorithms, and feature values in the training data. The purpose of normalization is to standardize all attributes to a consistent scale.

Apply Machine Learning

Here are the techniques to apply in Machine learning:-

- **KNN -** The algorithm learning algorithm that can tackle Tasks involving categorizing data into classes or predicting numerical values are respectively referred to as classification and regression problems. KNN adopts a lazy prediction approach and relies on the assumption that similar data points are situated near each other. By computing similarity measures, KNN groups new data and classifies them based on their similarities with existing records. The algorithm leverages a tree-like structure to measure the distance between data points. When making predictions when presented with a new data point, the algorithm identifies the K nearest neighbors in the training dataset, where K is a positive integer.
- **Random forest**, a popular machine learning algorithm coined by Leo Breiman and Adele Cutler, entails amalgamating the outcomes of numerous decision trees to generate a unified output. Given its versatility and user-friendliness, it is extensively employed to address classification and regression issues.
- **Support Vector Machine (SVM)** Refers to a learning technique that involves supervision partitions data into two distinct categories. It learns from a labeled dataset, and while it trains, it constructs the model. The aim of the SVM algorithm aims to determine the category or class that a novel data point belongs. This feature characterizes SVM as a non-binary linear classifier
- **Logistic regression** Logistic regression is a regression analysis that specializes in predicting the probability of a binary event. To understand logistic regression, it's crucial to first introduce the general concept of regression analysis. Regression analysis is a modeling technique used to establish the association between a dependent variable (usually labeled "Y") and one or more independent variables (usually labeled "X"). When multiple independent variables are utilized to predict or explain the outcome of the dependent variable, it is referred to

as multiple regression. Regression analysis can be applied for three main objectives: projecting the impacts of specific changes, predicting future values and trends, and evaluating the efficacy of different predictors.

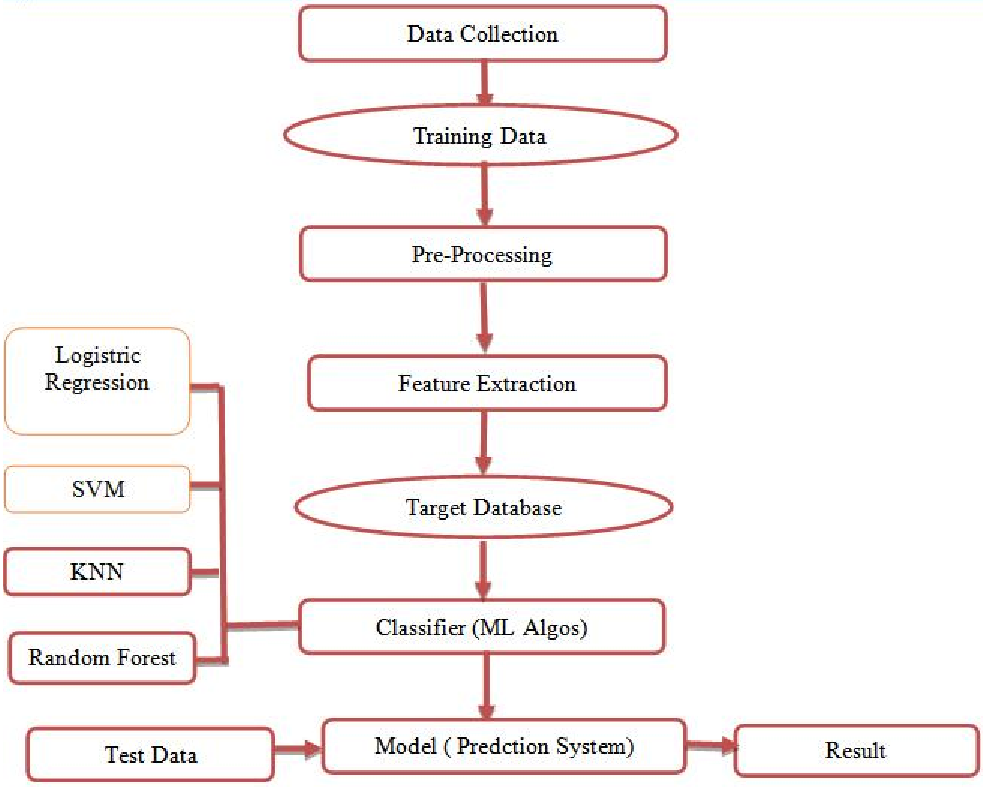


Figure 3: Diabetes prediction flow diagram

The first step involves importing the necessary libraries and loading the diabetes dataset. In step two, the data is pre-processed to eliminate missing values. Step three involves splitting the dataset into training and test sets using an 80-20 percentage split. Next, the machine learning algorithm we use four algorithm Like ML algorithms , is selected in step four. Step five involves building the classifier model using the training set. The classifier model is then tested using the test set in step six. In step seven, a comparing and evaluating the performance results of each classifier is carried out. Finally, in step eight, after analyzing the results The algorithm that performs the best is determined by evaluating various metrics..

IV. RESULT AND DISCUSSION

The table below displays the performance values of various classification algorithms, calculated using different measures. Based on the table, it is observed that Logistic regression exhibits the highest accuracy. Therefore, the Logistic regression machine learning classifier is capable of predicting the likelihood of diabetes with greater precision than other classifiers.

Classification Algorithm	Precision
Logistic regression(LR)	0.83
Support Vector Machine(SVM)	0.76
Random Forest(RF)	0.78
KNN	0.78

Table 2. Accuracy Measures

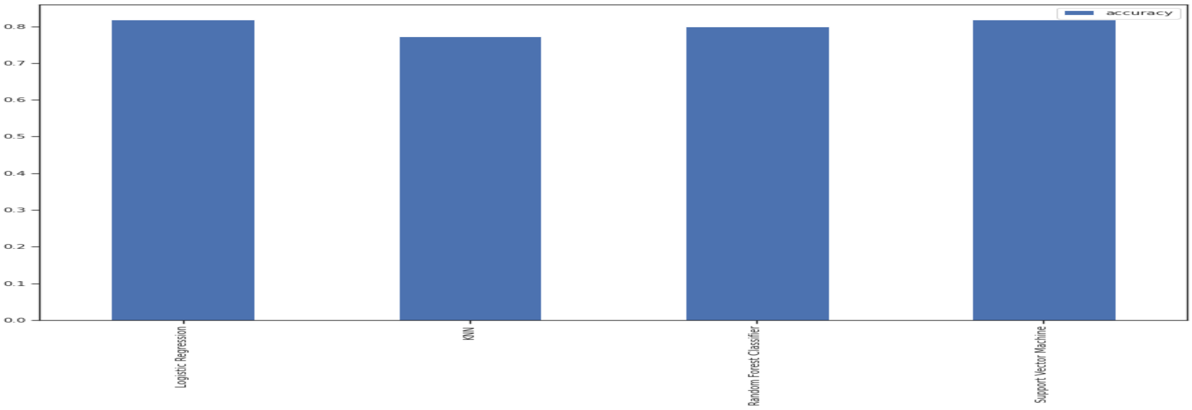


Figure 4: Results of the Accuracy achieved by machine learning techniques

V. CONCLUSION

The aim of this project was to assess the effectiveness or efficiency of Logistic Regression with other linear classifiers, Examples of such classifiers include SVM , KNN, and Random Forest(RF). The results of the comparison revealed that Logistic Regression outperformed all the other classifiers. The accuracy of Logistic Regression was found to be the highest, at **0.83%.** The proposed approach utilized ensemble learning and classification methods, which resulted in high accuracy levels. These experimental results can assist healthcare professionals by enabling early predictions and informed decisions, these classifiers can aid in the treatment of diabetes and potentially save human lives.

REFERENCES

[1]. Arwatki Chen Lyngdoh, Nurul Amin Choudhury, Soumen Moulik, “Diabetes Disease Prediction Using Machine Learning Algorithms”, IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) 2020, DOI: 10.1109/IECBES48179.2021.9398759 .

[2]. Mitushi Soni, Dr. Sunita Varma " Diabetes Prediction using Machine Learning Techniques" , Journal of Engineering Research & Technology (IJERT) 2020,

[3]. Sivaranjani S, Ananya S, Aravinth J, Karthika R, " Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction”,7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021 DOI: 10.1109/ICACCS51430.2021.9441935 .

[4]. Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal PatilB “Diabetes Prediction Using Different Machine Learning Approaches”IJSDR | Volume 7 Issue 5 , 2022.

[5]. Ashwini r, s m aiesha afshin, kavya v, deepthi raj“diabetes prediction using machine learning”ijrti | volume 7, issue 7 | issn: 2456-3315, 2022 .

[6]. G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8thInternational Conference on Reliability Infocom Technologies and Optimization (Trends and FutureDirections) (ICRITO), 2020 .

[7]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[8]. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[9]. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

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

[11]. M. F. Faruque, Asaduzzaman and I. H. Sarker, ”Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus”, 2019 International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 1-4, 2019.

[12]. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, ”Type 2 diabetes mellitus prediction model based on data mining”, Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.

[13]. Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of BigData ,6(1), p.13.(2019)

[14]. Sisodia,D. and Sisodia,DS,2018.Prediction of diabetes using classification algorithms. Procedia computer science,132, pp.1578-1585. (2018)

[15]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[16]. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[17]. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.