ISSN 2320-9984 02>
9 772320 998114
Volume 11 Issue 2 | June 2023
ISSN: 2320-9984 (Online)
International Journal of Modern Engineering & Management Research
Website: www.ijmemr.org

Analysis and Prediction of Gestational Diabetes Using Machine Learning Algorithm

Tanuja Sharma
M.Tech, Research Scholar
Computer Science and Engineering
Takshshila Institute of Engineering and Technology
Jabalpur (M.P.), India
Email: tanujacs13894@ gmail.com

Akshat Khaskalam
Assistant Professor
Department of Computer Science and Engineering
Takshshila Institute of Engineering and Technology
Jabalpur (M.P.), India
Email: akshatkhaskalam@takshshila.org

Abstract**—***Gestational diabetes is a common health condition affecting pregnant women, characterized by high blood sugar levels during pregnancy. Early identification and prediction of gestational diabetes can significantly contribute to the well-being of both the mother and the child. In this study, we employed a Voting Classifier machine learning algorithm, combining the Logistic Regression Classifier and Support Vector Classifier (SVC), to analyze and predict the presence of gestational diabetes using a dataset obtained from Kaggle. The dataset comprised records from 768 patients, with 268 patients diagnosed with diabetes and 500 patients without diabetes. By training the Voting Classifier on this dataset, we achieved an accuracy of 81.8% in predicting the presence or absence of gestational diabetes. The execution time for the model was 0.0338 seconds, indicating its efficiency for real-time applications. Furthermore, the area under the receiver operating characteristic curve (AUC-ROC) was calculated to assess the predictive performance of the model. The AUC-ROC value obtained was 0.78, suggesting a reasonably good discriminatory ability of the model in distinguishing between patients with gestational diabetes and those without. The results of this study demonstrate the potential of the Voting Classifier algorithm, combining Logistic Regression and SVC, for analyzing and predicting gestational diabetes. The high accuracy and efficiency of the model make it a promising tool for healthcare professionals in the early identification and intervention of gestational diabetes, ultimately contributing to improved maternal and fetal health outcomes. Further research and validation on larger datasets are recommended to enhance the robustness and generalizability of the proposed model.*

Keywords:— *Gestational Diabetes Prediction, Machine learning algorithm, Voting Classifier, Logistic Regression Classifier, SVC (Support Vector Classifier), AUC-ROC value.*

1. INTRODUCTION

The analysis and prediction of diabetes diseases using machine learning algorithms is an active area of research in the healthcare industry. With the increasing prevalence of diabetes worldwide, there is a growing need to develop accurate and efficient predictive models to identify individuals who are at risk of developing diabetes or who already have the disease. Machine learning algorithms are well-suited to address this challenge, as they can analyze large datasets and identify patterns that are difficult for humans to discern. Some popular machine learning algorithms that have been used for diabetes prediction include decision trees, random forests, support vector machines, and artificial

neural networks. In order to develop an accurate predictive model, researchers typically use datasets that include a variety of clinical and demographic variables, such as age, gender, body mass index, blood pressure, and blood glucose levels. By training the machine learning algorithm on these datasets, it can learn to identify patterns and make predictions about an individual's risk of developing diabetes. One of the key benefits of using machine learning algorithms for diabetes prediction is that they can be updated and refined over time as more data becomes available. This means that the accuracy of the predictive model can improve as more information is collected about the disease and its risk factors. Overall, the analysis and prediction of diabetes diseases using machine learning algorithms holds great promise for improving healthcare outcomes and reducing the burden of this chronic disease on individuals and healthcare systems around the world. The analysis and prediction of diabetes in pregnant women using machine learning algorithms is an area of research that has gained increasing attention in recent years. Gestational diabetes mellitus (GDM) is a form of diabetes that occurs during pregnancy and can have significant health implications for both the mother and child. Accurate identification of women at risk of developing GDM can help to improve maternal and fetal health outcomes. Machine learning algorithms can be used to analyze a range of clinical and demographic variables to predict the risk of GDM in pregnant women. Some of the key variables that are often considered include age, body mass index, family history of diabetes, previous history of GDM, and results of glucose tolerance tests. Various machine learning algorithms can be used for the analysis and prediction of GDM, including decision trees, random forests, logistic regression, and artificial neural networks. By training the algorithm on a dataset of pregnant women with and without GDM, the algorithm can learn to identify patterns that are predictive of GDM and generate accurate predictions for new patients. One of the challenges in developing predictive models for GDM is the need to balance accuracy with the complexity of the model. Models that are too complex may be overfit to the training data and not generalize well to new patients, while overly simplistic models may miss important predictive variables. Therefore, careful feature selection and model tuning are crucial to developing accurate and robust predictive models. Overall, the analysis and prediction of GDM using machine learning algorithms hold great promise for improving maternal and fetal health outcomes. By accurately identifying women at risk of developing GDM, healthcare providers can implement early interventions and improve health outcomes for both mother and child. There are three main types of diabetes: Type 1 diabetes, Type 2 diabetes, and gestational diabetes. Type 1 Diabetes: Type 1 diabetes, also known as insulin-dependent diabetes, is an autoimmune disease in which the immune system attacks and destroys the insulin-producing cells in the pancreas. This results in a lack of insulin in the body, which is necessary for glucose to enter cells and produce energy. Type 1 diabetes usually develops in childhood or adolescence, although it can also occur in adults. People with Type 1 diabetes require lifelong insulin therapy to manage their blood glucose levels. Type 2 Diabetes: Type 2 diabetes is a metabolic disorder characterized by insulin resistance, meaning the body is unable to use insulin effectively. It is the most common type of diabetes and is often associated with lifestyle factors such as physical inactivity, poor diet, and obesity. Type 2 diabetes can develop slowly over time and may be asymptomatic in its early stages. People with Type 2 diabetes may require lifestyle modifications, medication, or insulin therapy to manage their blood glucose levels. Gestational Diabetes: Gestational diabetes is a form of diabetes that develops during pregnancy. It is caused by hormonal changes that affect insulin sensitivity, resulting in high blood glucose levels. Gestational diabetes typically resolves after pregnancy, but women who develop gestational diabetes have an increased risk of developing Type 2 diabetes

later in life. Women with gestational diabetes require careful monitoring of their blood glucose levels and may require lifestyle modifications, medication, or insulin therapy to manage their diabetes during pregnancy. The main type of diabetes that can occur during pregnancy is gestational diabetes mellitus (GDM). GDM is a form of diabetes that develops during pregnancy and is caused by hormonal changes that affect insulin sensitivity. GDM typically develops in the second or third trimester of pregnancy and can cause high blood glucose levels, which can have significant health implications for both the mother and child. If left untreated, GDM can lead to complications such as macrosomia (a large baby), preterm birth, and an increased risk of cesarean delivery. After delivery, blood glucose levels typically return to normal, but women who have had GDM are at increased risk of developing Type 2 diabetes later in life. Therefore, it is important for women who have had GDM to undergo regular screening for Type 2 diabetes and adopt healthy lifestyle habits, such as regular exercise and a healthy diet. It is worth noting that some women may have pre-existing Type 1 or Type 2 diabetes before becoming pregnant. Women with pre-existing diabetes require close monitoring and management of their blood glucose levels during pregnancy to prevent complications for both the mother and child.

2. RELATED WORK

Voting Classifier -To Find Highest Model Accuracy

A voting classifier is a type of ensemble learning technique in which multiple machine learning models are trained and combined to make a prediction or classification. In a voting classifier, each model is trained on the same dataset using a different algorithm or with different hyperparameters. Once all the models have been trained, their predictions are combined using a simple majority vote, where the most common prediction is chosen as the final output. Voting classifiers can be used with both binary and multi-class classification problems. In a binary classification problem, the most common prediction is chosen as the final output, while in a multi-class classification problem, the prediction with the highest probability is chosen. The advantage of using a voting classifier is that it can improve the accuracy and robustness of a model, particularly if the individual models used in the ensemble are prone to different types of errors or biases. By combining the predictions of multiple models, a voting classifier can reduce the impact of individual model errors and provide a more accurate and reliable prediction.

Voting Classifier - To Find Least Execution Time of Model

Accuracy and execution time in each method is recorded in order to make better comparisons. Recording accuracy and execution time in different methods allows for better comparisons and evaluation of their performance. Accuracy refers to how close the results are to the true or expected values, while execution time measures how long it takes for a method or algorithm to complete its task. By recording accuracy, you can assess the effectiveness of different methods in achieving the desired outcome. Higher accuracy indicates better performance and reliability. Execution time is a measure of how long it takes for a method to process a given task or dataset. It provides insights into the efficiency and speed of a particular method. In general, shorter execution times are desirable, as they allow for quicker processing and analysis of data.

Voting: It refers to the process of combining the predictions of multiple models by taking a majority vote or averaging their results to make the final prediction. It can be used with both classification and regression problems.

Ensemble Classifier: Ensemble classifier, on the other hand, is a specific type of ensemble method that combines multiple individual classifiers to create a more powerful classifier.

3. PROPOSED WORK AND RESULTS

A. Proposed Work

1. Importing Libraries
2. Importing Data Set
3. Check Missing Values
4. Separate Features & Target Values
5. Diabetic Non-Diabetic Patient's Data Visualization
6. Data Standardization
7. Split the Dataset into Training & Testing Data Set
8. Model Creation & Calculating (Model Prediction / Accuracy Score / Execution Time / Confusion Metric)
   (i) Logistic Regression Classification
   (ii) Random Forest
   (iii) KNeighbors Classifier
   (iv) Decision Tree Classifier
   (v) Support Vector Classifier
   (vi) Naive Bayes Classifier
   (vii) Define the Voting Classifier
9. Comparison Between Different Model Accuracy Using Bar Chart / Pie Chart
10. Evaluating the Model
11. Calculate the AUC-ROC score to Evaluate Model Performance

B. Model Accuracy Comparison

- Using Bar Chart
- Using Pie Chart

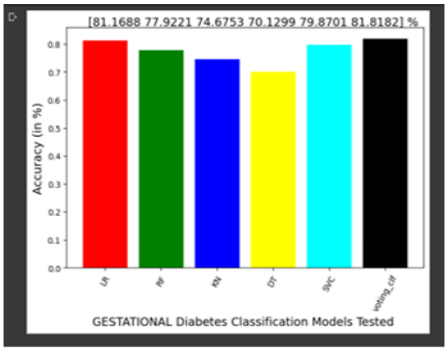

Figure 1: Model Accuracy Comparison Using Bar Chart

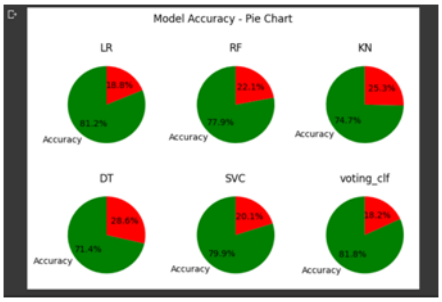

Figure 2: Model Accuracy Comparison Using Pie Chart

C. Area Under the Receiver Operating Characteristic Curve

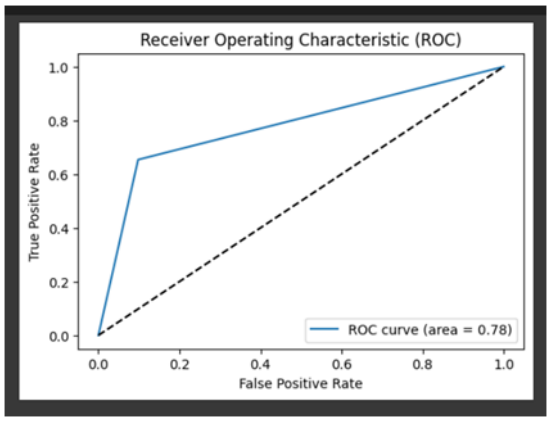

Figure 3: AUC-ROC Score to Evaluate Model Performance

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a numerical value that quantifies the performance of a classification model. It represents the area under the curve when plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds.

The AUC-ROC value ranges between 0 and 1, where:

AUC-ROC = 1: Indicates a perfect classifier that can perfectly separate positive and negative instances. It means the model has achieved a TPR of 1 (no false negatives) while maintaining an FPR of 0 (no false positives).

AUC-ROC = 0.5: Represents a random classifier with no predictive ability. The model's performance is no better than random chance. The ROC curve coincides with the

diagonal line connecting the bottom-left to the top-right corners of the graph.

AUC-ROC > 0.5 and < 1: Indicates the model's ability to discriminate between positive and negative instances. The higher the AUC-ROC value, the better the model's performance. A value of 0.7, for example, implies that the model has 70% discrimination power in distinguishing between positive and negative instances.

Table 1: Model Accuracy and Execution Time Comparison

Sr.No.	Existing Work	Algorithm	Accuracy in %	Execution Time (in Sec)
1.	Proceedings of the International Conference on Edge Computing and Applications (ICECAA 2022) IEEE Xplore Part Number: CFP22BV8-ART; ISBN: 978-1-6654-8232-5 978-1-6654-8232-5/22/$31.00 ©2022 IEEE	Decision Tree	82.7%	0.016
		Naive Bayes	83.5	0.013
		Logistic Regression	86.1	0.014
		Support Vector Machine	93.2	0.009
		ENSEMBLE (Naive Bayes, Support Vector Machine Classifier and Decision Tree)	94.5	0.010
2.	2021 International Conference on Information Technology (ICIT) Year: 2021 | Conference Paper | Publisher: IEEE Cited by: Papers (7)	Logistic Regression Classifier	80%	NA
		Linear Discriminant Analysis (LDA)	79%	NA
		Linear SVC	79%	NA
		Polynomial Kernal SVC	79%	NA
		Random Forest Classifier	82%	NA
		Voting Classifier (LDA, Logistic Regression Classifier & Random Forest Classifier)	80%	NA
3.	Proposed Model	Logistic Regression Classifier	81.2%	0.015574216842651367
		Random Forest Classifier	77.9%	2.089345932006836
		KNeighborsClassifier	74.7%	0.002199411392211914
		Decision Tree	71.4%	0.003590822219848633
		SVC Kernal Linear	79.9%	0.014202594757080078
		Voting Classifier (Logistic Regression Classifier & SVC)	81.8%	0.033846378326416016

4. CONCLUSION

The study successfully applied the Voting Classifier machine learning algorithm, consisting of the Logistic Regression Classifier and SVC, to analyze and predict gestational diabetes using the provided dataset. The Voting Classifier achieved an accuracy of 81.8%, indicating a reasonably good performance in classifying patients as having gestational diabetes or not. The AUC-ROC value of 0.78 suggests that the model has a moderate discriminatory power in distinguishing between positive and negative instances of gestational diabetes. Further improvements can be made to enhance the model's predictive performance. The Model's Execution Time of 0.033 Seconds is relatively low, indicating that it can make predictions efficiently, which is beneficial for real-time or time-sensitive applications.

REFERNCES:

[1] Nour Abdulhadi, Amjed Al-Mousa, "Diabetes Detection Using Machine Learning Classification Methods", 2021 International Conference on Information Technology (ICIT)

[2] C.S. Manikandababu, S. Indhu Lekha, "Prediction of Diabetes using Machine Learning", Proceedings of the International Conference on Edge Computing and Applications (ICECAA 2022). IEEE Xplore Part Number: CFP22BV8-ART; ISBN: 978-1-6654-8232-5.

[3] Srinivasa Rao Swarna, Sumati Boyapati, Pooja Dixit, Rashmi Agrawal "Diabetes prediction by using Big Data Tool and Machine Learning Approaches" Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020], IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3.

[4] Syifa Khairunnisa, Suyanto Suyanto, Prasti Eko Yunanto, "Removing Noise, Reducing dimension, and Weighting Distance to Enhance k-Nearest Neighbors for Diabetes Classification", 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-7281-8406-7/20/$31.00 ©2020 IEEE | DOI:10.1109/ISRITI51436.2020.9315515

[5] Irina-Oana Lixandru-Petre, "A Fuzzy System Approach for Diabetes Classification", The 8th IEEE International Conference on E-Health and Bioengineering - EHB 2020

[6] Rakesh S Raj, Sanjay D S, Dr. Kusuma M, Dr. S Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes",978-1-7281-0418-8/19/$31.00©2019IEEE.

[7] Priyanka Sonar, Prof. K. Jaya Malini, "Diabetes Prediction Using Different Machine Learning Approaches", Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4.

[8] Sara A. Aboalnaser, Hanan R. Almohammadi, "Comprehensive Study of Diabetes Miletus Prediction using Different Classification Algorithms", 978-1-7281-3021-7/19/$31.00 ©2019 IEEE | DOI 10.1109/DeSE.2019.00033

[9] Melky Radja, Andi Wahju Rahardjo Emanuel, "Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction", 2019 5th International Conference on Science in Information Technology (ICSITech), 978-1-7281-2380-6/19/

S31.00 ©2019 IEEE.

[10] Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", 2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018.

[11] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, " Analysing Feature Importances for Diabetes Prediction using Machine Learning", 978-1-5386-7266-2/18/$31.00 ©2018 IEEE

[12] Ayman Mir, Sudhir N. Dhage, "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare", 978-1-5386-5257-2/18/ $31.00 c 2018 IEEE.

[13] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018.

[14] Wenqian Chen, Shuyu Chen, Hancui Zhang, Tianshu Wu, "A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree", 978-1-5386-0497-7/1/$31.00 ©2017 IEEE

[15] M. Durgadevi, Dr. R. Kalpana, "Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes", 2017 Ninth International Conference on Advanced Computing (ICoAC), 978-1-5386-4349-5/17/$31.00 ©2017 IEEE

[16] Girdhar Gopal Ladha, Ravi Kumar Singh Pippal, "A review and analysis on data mining methods to predict diabetes", 2017 7th International Conference on Communication Systems and Network Technologies, 978-1-5386-1860-8/17/$31.00 ©2017 IEEE, DOI 10.1109/CSNT.2017.64.