Diabetes Disease Prediction and Customized Diet Recommendations using Machine Learning model

*Senbagavalli M[1], *Shashi Kant Gupta*[2], *Saswati Debnath*[3]
*[1,3]Post Doctoral Researcher, LinColn University College, Malaysia; [2]Adjunct professor, LinColn University College, Malaysia.
[1]pdf.senbagavalli@lincoln.edu.my, senba1983@gmail.com ; [2]shashigupta@lincoln.edu.my, [3]pdf.saswatidebnath@lincoln.edu.my

ABSTRACT
Diabetes is a chronic metabolic disease that affects millions of people worldwide. Early detection and precise diabetes prediction are critical for managing diabetes effectively and preventing complications. In this study, we propose a machine learning-based approach to diabetes disease prediction. The objective is to develop a precise model that can determine a patient's risk of developing diabetes based on a range of medical traits. We collected clinical and demographic data from a diverse group of individuals, including age, BMI, blood pressure, blood sugar levels, and family history. The dataset was preprocessed to address feature scaling, outliers, and missing values. Several machine learning methods were applied, such as support vector machines (SVM), logistic regression, and random forest. After each algorithm was trained on a subset of the dataset, its performance was evaluated using k-fold cross-validation. The models were compared using metrics for accuracy, precision, recall, and F1-score. Our results demonstrated that the linear regression approach had the best prediction accuracy, averaging 85% across all folds. The confusion matrix analysis demonstrated a significant reduction in false negatives, indicating the model's ability to accurately identify individuals at risk for diabetes. The importance of the selected variables—such as BMI and glucose levels—in the diagnosis and prediction of diabetes is emphasized. This work advances the field of predicting the onset of diabetes disease by employing machine learning techniques to produce a reliable and accurate model. The proposed approach may assist healthcare professionals in identifying individuals at risk for diabetes at an early stage, allowing for timely interventions and customized treatment plans. To improve the model's generalizability and efficacy in real-world clinical scenarios, further validation on larger and more diverse datasets is recommended.

Keywords:
Diabetes Prediction, Machine Learning, AI Diagnosis, Personalized Diet, Health Management, Predictive Analytics

1.INTRODUCTION

The data in the medical industry is huge. Information in these databases may be organized, semi-structured, or unstructured. Big data analytics is a way to analyze large volumes of data and discover hidden patterns and information to extract insights from data. According to the current situation, diabetes mellitus (DM) has developed into a severe illness. in developing countries like India. Many people suffer from diabetes mellitus (DM), which is classified as a non-communicable disease (NCB). According to 2017 data, 425 million people worldwide have diabetes. Diabetes kills 2 to 5 million people each year. This number is estimated to reach 629 million by 2045. Insulin Dependent Diabetes (IDDM) is a classification of **Type 1** Diabetes Mellitus (DM). In this type of diabetes, patients need insulin injections because their bodies cannot produce enough insulin on their own. **Type 2** Non-insulin dependent diabetes mellitus, or type 2, is a common abbreviation. The body's cells utilize insulin incorrectly in this kind of diabetes, which causes its symptoms. When pregnant women who have not yet developed diabetes have high blood sugar levels. **Type 3** gestational diabetes develops. Diabetes is linked to long-term consequences. Patients with diabetes are at risk for a wide range of medical issues. Combining several machine learning algorithms, data mining strategies, and statistical tools,

predictive analytics is a strategy for investigating data and forecasting future occurrences. By analyzing the estimations of clinical data, significant judgements and predictions may be made.

Regression and machine learning methods are both applicable to predictive analysis. Using predictive analytics, improve health outcomes by improving patient outcomes, improving resource utilization, and making better diagnoses. Machine learning, one of the most important aspects of artificial intelligence, leads to the creation of computer systems that can learn from the past without any programming whatsoever.
The current state of affairs makes machine learning an urgent necessity to replace human labor by automating tasks that are prone to mistakes. Diabetes is presently diagnosed through tests like oral blood pressure and fasting blood glucose. This approach, though, requires time. The development of a diabetes assessment model utilizing machine learning algorithms and data mining is the main goal of this project. The format of this document is as follows: Part II gives a review of earlier studies on using machine learning approaches for early diabetes diagnosis and categorization. Chapter III outlines the goal of overcoming this issue. Chapter IV offers suggested diabetic testing guidelines. In Chapter V, experimental findings are provided, followed by conclusions and suggestions.

2. LITREATURE SURVEY

[1]They examined the prevalence of end-stage renal disease (ESRD) and the correlation between education and coronary heart disease (CHD) and ESRD in individuals in Norway who had type 1 diabetes diagnosed as a kid.[2] Processed to assess the viability of the 5:2 diet and its effects on weight, metabolic and hormonal risk indicators for cardiovascular disease (CVD), and glycemic management in overweight and obese patients with and without type 2 diabetes (T2D). They know if levels of standard capillary blood ketone tests might forecast a person's risk of developing DKA in the future. [3] Founded that using a network pharmacological approach and experimental assessment, it will be possible to confirm the effectiveness of FFSLD in treating T2DM and investigate probable mechanisms of action.[4] This Study aimed to when immunosuppressant medications like tacrolimus are offered to transplant recipients, a side effect known as post-transplant diabetes mellitus (PTDM) might occur. The goal of this study was to look at the PTDM risk variables and develop a model for predicting PTDM risk.
[5] population made up of people having a history of heart failure (HF), multiple risk factors for heart disease (MRF), or established atherosclerotic cardiovascular disease (ASCVD). This meta-analysis was carried out to demonstrate various advantages of SGLT-2is on the renal composite based on baseline ASCVD or MRF and HF since they represent two separate entities from a molecular, pathophysiological, and therapeutic standpoint. [6] This study may be divided into two main sections: the first portion assessed the amounts of trace elements and their relationships to clinical parameters according to sex; the second section built auxiliary prediction models that incorporated the trace elements and clinical factors. depicted the precise research design of this work, and the following provided a thorough explanation of the methodology. [7] Study on Precision medicine depends on accurate big data analysis. While statistical analysis frequently reveals a large number of connections, relationships by themselves lack predictive power. Clinicians, scientists, and ultimately patients suffer when association and prediction are confused. We looked for diabetes-related studies that had the word "prediction" in the title. We evaluated if these publications have metrics that are useful for prediction. [8] Given that gestational diabetes mellitus (GDM) increases the likelihood of bad pregnancy outcomes, it would be possible to give preventive and therapeutic therapies to high-risk women while avoiding the needless treatment of low-risk women. Our goal was to create, validate, and assess the therapeutic value of a prediction model for unfavorable pregnancy outcomes in GDM-affected women.

[9] Proposed the ascertain the prevalence of end-stage renal disease (ESRD) and the correlations between education and coronary heart disease (CHD) and ESRD in individuals in Norway

who were diagnosed with type 1 diabetes of childhood and who were subsequently followed up on. [10] Used the decision tree, random forests, and neural network to predict diabetes. The data was gathered from the Chinese Liuzhou physical examinations. The dataset's dimension was reduced using the PCA. To check the method's general applicability, they chose several ML techniques and carried out independent tests on them. [11] proposed a hybrid methodology to identify type 2 diabetes. K-mean clustering is used to extract unknown, hidden properties from the dataset, and then a Random Forest and xg Boost classifier are run on top of it to get more precise results.[12] Predicted diabetes employing the neural network, random forests, and decision tree. The dataset was collected from physical examinations in Fuzhou, China. The PCA was used to reduce the dimension of the dataset. They employed a few ML approaches for independent experiments to verify the method's wide applicability.

[13] All of the qualities were categorized using NBs. Next, NBs used GA as an attribute selector to classify the data based on the chosen attributes. The study results highlight the effectiveness of this work on PIDD and provide a more precise diagnostic classification. Joshi and Chowhan (2018) specifically used three supervised machine learning methods. [14] Study on gestational diabetes mellitus (GDM) increases the possibility of unfavourable pregnancy outcomes, it would be conceivable to provide high-risk women with preventative and curative medicines while avoiding the unnecessary treatment of low-risk women. A prediction model for unfavourable pregnancy outcomes in women with GDM was something we wanted to develop, validate, and evaluate for its therapeutic potential.

3.PROPOSED METHODOLOGY

Create a machine learning system that uses patient demographics, medical history, and biomarkers to predict the chance of getting diabetes with accuracy using methods like random forest classification or logistic regression. To guarantee robustness and generalizability, verify the model's performance using cross-validation techniques after training it on a large dataset of diabetes patients and non-diabetics while optimizing for high accuracy, sensitivity, and specificity.

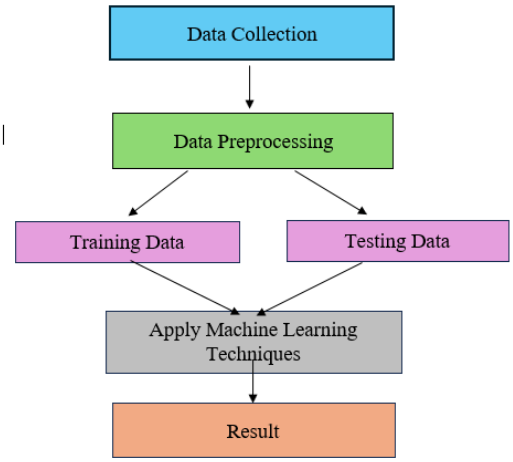

Figure - 1 Architecture of Diabetes Prediction Model

Figure 1 for diabetes prediction model. This model has five different modules. These modules include

i. Dataset Collection
ii. Data Pre-processing
iii. Training and Testing Data Set
iv. Build Model
v. Evaluation

Let’s have a look at each model briefly

3.1. Dataset Collection

These models use data gathering and analysis to spot patterns and trends that aid in result prediction and evaluation. The dataset is described in the section below. 1200 data for this diabetes include 9 attributes.

Table 1. Dataset Information

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Job Type(Office-work/Field-work/Machine-work)	No
Outcome	C

.

3.2. Data Preprocessing

Overlapping data is processed at this stage of the model to establish accuracy and accuracy. This information is missing. Therefore, we attributed missing values exist for several chosen characteristics, including age, blood pressure, skin thickness, glucose level, and BMI, as these variables cannot have a value of 0.

The dataset should then be scaled to normalize each value.

3.3. Comparative Analysis

Various machine learning methods can be used to predict diabetes such as support vector machines, decision trees, random forests, and logistic regression. The scope and depth of the dataset, the characteristics used as input, and the level of accuracy and interpretation required affect the selection process.

- **Logistic Regression**

Logistic regression is a popular algorithm in machine learning for predicting binary outcomes. It works by estimating the probability of an event occurring based on different strategies.

- **Random Forest**

Multiple decision trees are combined in Random Forest to improve accuracy and lessen competition. Random forests can be used to create more accurate diabetes models by combining predictions from multiple decision trees based on a variety of different input strategies.

- **Support vector machines (SVMs)**

Support vector machines are supervised learning algorithms that may be used for replication or classification. Support vector machines can be used to predict diabetes to determine the best hyperplanes that separate diabetics from non-diabetics based on parameters such as age, BMI and glucose levels.

3.4. Train and Test Datasets:

Divide a dataset into train and test sets so that you may train a model on a portion of the data and assess how well it performs when given new data.

Use relevant measures to assess the model's performance, including the F1 score, accuracy, precision, recall, and area under the ROC curve (AUC-ROC). Perform validation or use techniques such as layered sampling to get a good estimate of sample performance.

4. Results and Discussion

Several machine learning methods have been applied to the data. After comparison, the logistic

regression (82%) algorithm achieved the highest accuracy when compared to other algorithms. The results are as follows in table 1.

Table 1: Comparative analysis using different machine learning models

Algorithms	Accuracy
Random Forest	79
Logistic Regression	82
SVM	81
Decision Tree	78
XG Boost	79
KNN	76

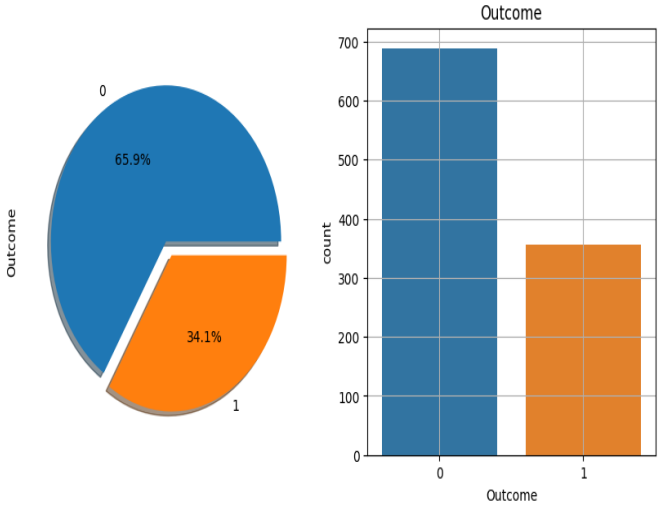

Figure - 2 Outcome of Positive and Negative

Positive (1): 356
Negative (0): 6

Histogram Analysis

Histogram analysis enables users to examine the distribution and characteristics of data, histogram analysis is a useful method for information analysis in the context of diabetic disease prediction using machine learning algorithms. This study investigates the benefits, techniques, and potential impact on improving prediction models of the use of histogram analysis in diabetes prediction. The study explores the potential applications of histograms to observe feature distributions, identify patterns, and understand the relationship between attributes and the target variable.

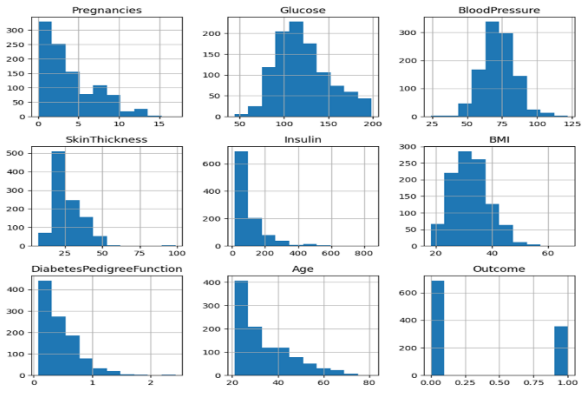

Figure – 3 Histogram Analysis Outcome

Pair Plot

Pair plot analysis is a comprehensive method for concurrently visualizing the correlations between many variables, making it a useful tool when studying datasets for the goal of predicting the occurrence of diabetes sickness using machine learning algorithms. This study will outline the benefits, technique, and potential impact on prediction model improvement of pair plot analysis in order to better understand how it may be utilized to predict diabetes. In order to provide valuable insights for feature engineering, model development, and feature selection, the paper explores the use of pair plots to identify interactions, correlations, and patterns between features.

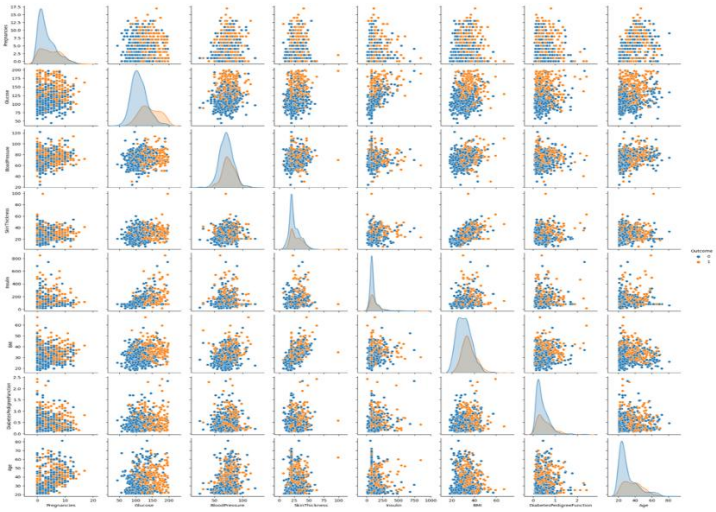

Figure - 4 Pair plot Analysis Outcome

We connected the flask, a backend web development framework, and frontend we used as html and CSS, in the first page we created a form for diabetes patient and we are checking whether they have diabetes or not so we giving their health rating for blood pressure, glucose, skin thickness, BMI, pregnancy, insulin, and diabetes pedigree function this all the rating we will enter in this form after entering click on submit button then it will display the results of our analysis. The second page will demonstrate whether a person has light or moderate diabetes and how the calculation is made using the patient's information from the first page alone. It displays the patient's level of diabetes and, after pressing the submit button, advances to the next page, which contains the patient's meal plan. Here, we outline what the patient wants to eat for breakfast, lunch, dinner, and snacks so that they can better manage their diabetes. We also specify how much protein, carbohydrates, fat, and fiber they should consume in each meal. The main goal of this project is to provide meal plans for people with diabetes so they can better manage their condition. To produce a diabetes prediction, we must fill out a form on the first page of the output, which includes the user's age, blood sugar level, body mass index (BMI), and other parameters.

Figure – 5 Home page of the proposed system

We were able to determine the person's type of diabetes on the second page of the result. Where we were able to offer a food regimen tailored to his kind of diabetes.

Figure - 6 Diabetic Identification

We may create a diet plan for a patient on the third page of the Output based on their type of diabetes, and we can assess the patient's intake of carbohydrates, protein, fat, and fiber at various intervals.

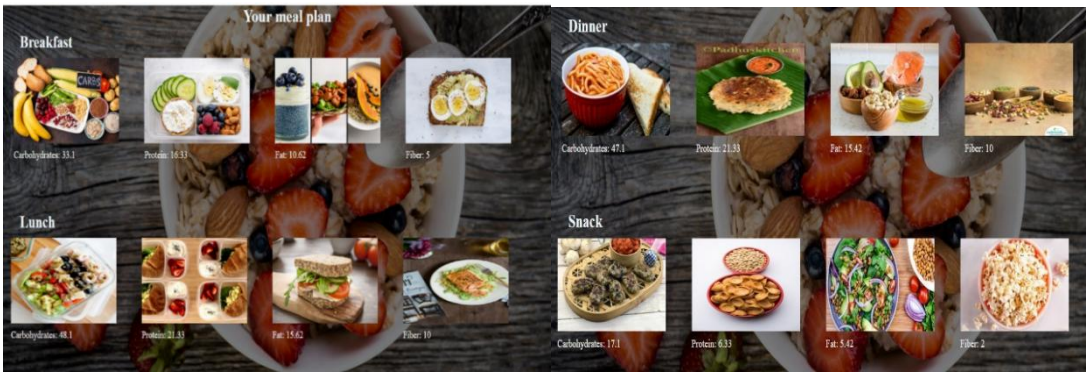

Figure – 7 Diet Recommendations

6.CONCLUSION

The logistic regression (81%) has provided the highest accuracy, using different learning machines for datasets and different algorithms for classification. From the data, we observed a comparison of the accuracy of machine learning algorithms. Using this information is not prior knowledge to enable the model to measure blood sugar accurately and accurately. This research could be extended to determine the likelihood that non-diabetics will develop diabetes in the next few years.

7.REFERENCES

[1]. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar, “Predictive Analytics in Diabetes Patient Data Using Machine Learning and Hadoop”, I-SMAC International Conference, 978-1-5090-3243-3, 2017.

[2]. Ayush Anand thiab Divya Shakti, "Diabetes Prediction Based on Individual Lifestyle Indicators," 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3]. B. Nithya and Dr..V. Ilango, “Predictive Analytics in Healthcare Using Machine Learning Tools and Techniques,” International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7, 2017.

[4]. Saravana gambling N M, Eswari T , Sampath P and Lavanya S Drs, "Predictive methods for diabetes data analysis in big data", 2.International Big Data and Cloud Computing Symposium, 2015.444 5.

[5].Aiswarya Iyer, S. Jeyalatha, thiab Ronak Sumbaly, "Diagnosting Diabetes using Classification Mining Techniques", International Journal.Data Mining and Knowledge Management Processes (IJDKP), Cilt 5, Sayı 1, Ocak 2015.444 6.

[6]. P. Suresh Kumar thiab S. Pranavi "Performance Analysis of Machine Learning Algorithms in Diabetes Dataset Using Big Data Analytics", Infocom International Conference on Technology and Unmanned Systems, 978-1-5386-0514-1, 18 Kaum Ob Hlis Ntuj txog 20, 2017

[7]. Mani Butwall thiab Shraddha Kumar, "A Data Mining Approach to Diagnosing Diabetes Using Random Forest Classifiers",International Journal of Computer Applications, Vol. 120 - No. 8, 2015.

[8]. K. Rajesh thiab V. Sangeetha, "Data Mining Methods and Techniques for Diagnosing Diabetes", International Journal of Engineering and Innovative Technology (IJEIT), Vol. 2, Issue 3, September 2012.

[9]. Humar Kahramanlı and Novruz Allahverdi, "Building a hybrid system for diabetes and heart disease", Expert System and Article: International Journal, Vol. 35, no. 1-2, July 2008.

[10]. UN.Patil, R.C. Joshi and Durga Toshniwal, "Integrated Rules for the Classification for the Type 2 Diabetes", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, 9-11 February 2010

[11]. Dost Muhammad Khan1, Nawaz Mohamudally2, "Integrating K-Means and Decision Trees (ID3) into More Efficient Data Mining Algorithms", Journal of Computing, Vol. 3, no. December 12, 2011.

[12] Diagnosis and follow-up of children with type 1 diabetes in Norway from 1973 to 2017 Insulin auto-antibodies population.

[13] International Journal of Computational Intelligence in Engineering 2021 Hengil Zhao Diabetes prediction based on machine learning and smart network application development.

[14] Journal of Ethnopharmacology 2022 Yilin Wei, Honngli Mu Compound Huishi Liufang alleviates type of insulin resistance: a communication discussion Research and Research Review.

[15] Research on Diabetes Research and Treatment 2022 Xianjing HU, Genting Dong, Xin Liu Type 2 Diabetes in Cardiovascular Risk Patients: A Systematic Review of Prediction Models 4442 Changib 1 Three Type 1 Diabetes.

[16] Reference Enhanced Evidence 2019 Javier Zamora, Helensa J. Teede of retina Deep Learning Images Enhance Known Risk Factors for Prediction of Diabetic Cardiovascular Disease.

[17] 2018 from Biomedical Computing Methods and Procedures and a Machine Learning Perspective.

[18] Science Direct 2021 John Allotey, Holley Wang Comparison of Machine Learning Algorithms for Diabetes Prediction.
[19] Health Reviews 2020 Borja M. Femendez-Felix Predicting Type 2 Diabetes Profile Using Genome-Wide Polygenic Risk Score Machine Learning Analysis of 10-Year Population-Based Prospective Cohort Study