Type-2 Diabetes Prediction in Women Using Machine Learning in Malabar Area

SruthyK.G.
AssistantProfessor,dept.ComputerScience
MEA Engineering College
Perinthalmanna, India
sruthydas.abhi@gmail.com

MuhamedRafsalAmeenM
dept.ComputerScienceandEngg.
MEAEngineeringCollege
Perinthalmanna, India
rafsalpkd@gmail.com

MuhammedFadilKS
dept.ComputerScienceandEngg.
MEA Engineering College
Perinthalmanna, India
muhammedfadil030@gmail.com

MuhammedShahim
dept. Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
shahimellathodi@gmail.com

Muhammed Shamsheer Shajahan
dept. Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
shamsheercp8921@gmail.com

***Abstract*—Forahealthylife,ourmetabolismshouldbeworking properly. Also, the physical, mental, and social well-being of a personareimportant.Thereisawiderangeofdiseasesthat may have severe impacts on our health. Diabetes is one of them. Glucose is essential for the proper functioning of our body. Inthe case of Type2 diabetes, there is presence of insulin, but the cells do not respond to it. So the glucose will stick in the blood, thereby increasing the blood sugar level. This condition is more complicated in women. It increases the risk of heart disease, blindness, kidney disease, depression, etc. Here we introduce a system to predict the probability of a person(a woman) being affected by Type2 diabetes. To get a better result, we used Machine learning algorithms like Decision Tree, Support Vector Machine, and Logistic Regression.**

***Index Terms*—Machine Learning, Classification Algorithms, Support Vector Machine, Decision Tree, Logistic Regression.**

I. INTRODUCTION

Diabetesisagrowingglobalhealthissue,andType2 diabetes is the most common form of the disease. It is char-acterized by high levels of glucose in the blood and is oftenassociatedwithlifestylefactorssuchaspoordiet,physical inactivity,andobesity.TheMalabarareaisnoexceptionto this trend, with increasing rates of Type 2 diabetes among its population. Early detection and prevention of Type 2 diabetes are crucial for reducing its impact on public health. However, thetraditionalmethodofdiagnosingType2diabetes,which reliesonmanualassessmentsofdemographicandlifestyle factors,andoneofthemaindrawbackisthatittakesmore timeandalsothereisachanceoferrorcausedduetohuman.

Inlightofthis,theobjectiveofthestudyistobuild a model which enables prediction of Type 2 diabetes in women in the Malabar area which depends if demographicand lifestyle factors . The study will use a large and diverse dataset of womenin the Malabar area to trainand evaluate the model. This study is significant because it has the potential to improvetheaccuracyofType2diabetespredictions,leadingto earlierdetectionandprevention.Additionally,itcontributesto the field of machine learning by demonstrating the potential for its application in healthcare, specifically in the area of diabetes prediction.

The study mainly aims to provide basis for future research in this area, including the development of more accurate and sophisticatedmodelsforType2diabetesprediction.Thestudy willalsohavepracticalimplicationsforhealthcareprofession- als and policy makers, as it can inform the development of effective strategies to prevent and manage Type2 diabetes in the Malabar area.

II. RELATEDWORKS

The primary objective of the research was to develop and implement a machine learning approach for predicting dia-betes. The achieved classification accuracy of 56 is promising and can help healthcare providers make early predictions and decisionstotreatdiabetesandsavelives.Movingforward,the researchers intend to replicate their analysis of classification models using sophisticated machine learning algorithms and real world datasets. Machine learning, which is a technology that enables machines to automatically detect patterns using various techniques and data, can be leveraged for future decision-making. It offers several algorithms that allow ma-chines to comprehend current events and make appropriate decisionsbasedonthatunderstanding.LogisticRegression is a widely used machine learning approach for binary clas-sificationthatprovidesadiscretebinarynumberbetween0 and 1, representing two possible outcomes. The strategic function, was developed by analysts to describe population growth patterns in biology and other fields, and it increases rapidly and peaks near the carrying limit of the environment. Logistic Regression is a simple yet effective algorithm thatcan be used as a performance baseline for most tasks. [1].

A separator was used to predict diabetes in this study. Its efficiency was improved through techniques such as outlier rejectionandfeaturebuilding.Insulin,glucose,andskin

stiffness are important factors in diabetes according to thedata analyzed. The Diabetes Pedigree Factor also impacts the prognosis. The ML model's performance can be improved by tweaking its parameters. A cloud based interface is used bythe model to predict diseases [2].

Machine Learning can improve diabetes diagnosis and treatment through classification techniques such as Logistic Regression, K-Nearest Neighbors, SVM, and Random Forest [14].Thegoalistodevelopanaccuratepredictiontool using these methods and analyze its performance. The current accuracy rate is less than 70, so the authors suggest using associative methods, which combine multiple techniques for higher accuracy [3].

The project successfully designed and implemented a Diabetes Prediction system using various machine learning methods including SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting classifiers [14]. The system achieved 77 classification accuracy and can assist healthcareinearlypredictionandtreatmentofdiabetestosave lives [4].

A new Machine Learning-based decision support system wasproposedinthisstudy.Thesystemutilizesfuzzylogic to integrate two commonly used machine learning techniques. The proposed model consists of two main phases. In the Training, data is acquired, preprocessed, and classified using SVMs and ANNs. Various accuracy measures are used to evaluate the performance of the model. The outputs of the SVM and ANN models are then fused using fuzzy rules to makeafinalprediction.Thefusedmodelisstoredinthecloud. In the Testing, the preprocessed training model is loaded from thecloudandusedtopredictwhetheradiabetesdiagnosis is positive or negative. The ANN model is trained using a preprocessedtrainingdatasetwithBayesianregularizationand 16 hidden layers between input and output neurons [5].

The paper describes the creation and evaluation of various Support Vector Machine (SVM) models using different types of kernels such as linear, sigmoid, polynomial, and radial basis. The performance of these models has been assessed,and it was found that the linear kernel performed better in predictingthediseasewhencomparedtotheotherkernels.The dataset used in the study is the Pima Indian diabetes dataset, and normalization was applied before and after comparing the different kernels with SVM. In the case of linear data, the linear kernel is a commonly used kernel with a regularization parameter (e=1.0) to speed up execution. On the other hand, radial basis kernels are more expensive than linear kernelsand are preferred for non-linear data. The Y parameter is used in this kernel, and as its value increases, the model becomes overfit, while decreasing its value results in underfitting. The Polynomial kernel is complex to compute and non-linear in nature [6].

Inthisresearch,variousMachineLearningandDeepLearn- ing algorithms were compared for predicting diabetes. The studyfoundthatRFalgorithmhadthehighestaccuracyof 83.67inpredictingdiabetes,whileSVMhadanaccuracyof 65.38andDLhadanaccuracyof76.81onthedataset.Inthe future, the researchers plan to improve the feature extraction step by using an automatic deep feature extraction approach. For all experiments, 60 of the data was used for training and validation, and 40 for testing. The performance of the models was evaluated using various metrics such as overall accuracy, Kappa Coefficient, precision, recall, and f-measure. The study was repeated ten times to avoid any bias in the models. The CNNwasusedforfeaturerepresentationoftheinputdata,and it was found that it reduced the complexity of the network by applying the convolution [7].

The study employed several Machine Learning algorithms to classify a dataset.. A pipeline was used to apply AdaBoost. The research compared algorithm accuracies on two different datasets,revealingthattheproposedmodelimprovedtheaccu- racy and precision of diabetes prediction compared to existing methods. Future research could explore the likelihood of non-diabetic individuals developing diabetes in the coming years. Machine learning is a crucial aspect of artificial intelligence that enables computers to learn from past experiences without requiring individual programming for each case. It is essential in automating processes and minimizing errors. While the existing diabetes detection method involves lab tests that are time-consuming, the study aimed to create a predictive model for diabetes prediction using Machine Learning Algorithms and Data Mining techniques [8].

In This paper Machine Learning plays a critical role in predicting diabetes. The primary objective of executing a dataset is to predict diabetes based on certain diagnostic measurements. The data contains numerous attributes as risk factors such as BMI, age, and blood pressure. These risk components play a significant role in predicting diabetes. In the pre-processing stage, the original dataset is split into two subsets: diabetic and non-diabetic. The pre-processed dataset isthenfusedtocreateanoveldataset.Featureselectionin machine learning involves removing redundant attributesto choose the best feature. This improves the classifier's simplicity and provides more scalability, stability, and accu- racy in optimal feature selection. RF-WFS is an algorithmthat employs a resampling technique to create various data subgroups.Thefullfeaturesignificancemeasurevaluesof the data subset are inserted together for acquiring the final feature rank measure value, and the allocation of rank is performed on the features depending upon their significance. The optimal feature subgroup is provided through sequential backward selection [9].

The study employed six different Machine Learning methods to classify data collected from online and offline questionnairesrelatedtodiabetes.Theresultswerecomparedusing variousstatisticalmeasures,andthesamealgorithmswerealso applied to the PIMA database. All the models produced good results for parameters such as precision, recall sensitivity, etc. Thestudyrevealedthatvariablessuchasage,familyhistoryof diabetes, physical activity, regular medication, and gestational diabetes had the greatest significance in predicting diabetes. These findings may be useful for predicting other diseases,andthereispotentialforfurtherresearchandimprovemen tby

incorporatingotherMachineLearningalgorithms[10][12].

III. METHODOLOGY

We train the machine with the dataset that we have and test it with the Machine Learning algorithms like Support Vector Machine, Logistic Regression and Decision Tree [13]. Andwe will deploy our model by choosing the algorithm which give the highest accuracy. In Data Preprocessing health care values consists of so many impurities which may affect the efficiency of data.The aim of the data set is to make sure whether the person is diabetic or not. The data is splitted into test data and another one is training data. And then the model is trained with the help of training data. Test data is used to check whether the trained model gives correct output or not. Classification algorithm are mainly used in case of discrete data. The main purpose of classification algorithm is to fix the decision boundary so that it enables splitting of dataset into different classes. Classification algorithms are distinguished into binary classifier and multi class classifier. After training the model the accuracy of the algorithm is checked. The algorithm which shows highest accuracy is selected and using that algorithm the model is deployed.

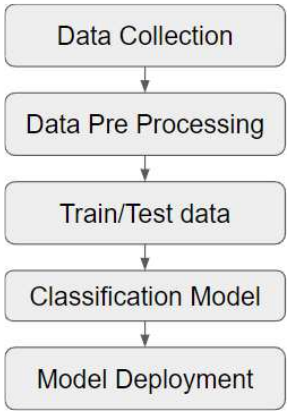

Fig.1.SystemProcessingFlowchart

A. DataCollection

The data collection process for this project involved gathering relevant data from various sources. The variables that were essential for the prediction model were identified and defined, such as age, body mass index, heart rate, family history, lifestyle factors, and medical history. This data was collected through medical records, and physical examinations of the participants. We also seek help from pima data set where we got real time data. We also conducted surveys by creating a google form distributed it among all peoples we know through social media and mainly through personal chat. Oncewecollecteddataintogooglesheetsandbegananalyzing it.Wewereabletoaccuratelypredictaftercarefullyexamining the data. It was crucial to ensure that the data was accurate, reliable, and representative of the target population to prevent biasandimprovethepredictivepowerofthemodel.The data collection process was carefully planned and executed to achieve this goal. The collected data was evaluated to ensureit met the required standards. The collected data was preprocessed to make it suitable for analysis and modeling. This involvedcleaning,transforming,andorganizingthedata,such as removing duplicates, filling missing values, normalizing, andencodingcategoricalvariables.Thedatacollectionprocess wasacriticalcomponentofthisproject.Byfollowingacareful and diligent approach, accurate and reliable data was gathered to create a successful prediction model for this project.

SL.NO	Attributes
1	Pregnancy
2	Calorie Consumption Per Day
3	Heart Rate
4	Glucose
5	Insulin
6	Body Mass Index
7	Diabetes pedigree
8	Age

Table.1Attributes

B. DataPreprocessing

Datapreprocessingisacrucialstep.Inthisproject,wehave collected data from multiple sources, including surveys, the Pima Indian dataset, and medical records. The first step indata preprocessing is to ensure that all the data is in the same format and is consistent. This includes converting data types, dealing with missing values, and identifying outliers.

After we have cleaned and standardized the data, the next stepistoprepareitforanalysis.Thisincludessplittingthedata intotrainingandtestingsets,aswellasfeatureengineering.In our project, we may want to engineer features that are specific towomen,suchasreproductivehealthdata.Wemayalsowant to normalize or scale the data to ensure that all features are on the same scale. Overall, the data preprocessing step is critical in ensuring that our machine learning model is accurate and reliable. By completing these steps, we can ensure that our machine learning model is accurate and can provide useful predictions for healthcare professionals.

In this project, we need to ensure that the data is of high qualityandisrepresentativeofthepopulationwearestudying. This requires us to carefully examine the data for errors, outliers, and missing values. We may also need to perform data imputation, which involves filling in missing data with estimated values based on statistical methods.

C. ClassificationAlgorithm

1) Decision Tree: Decision Tree is used in case of both classification and regression problems but mainly it is used in caseofclassificationproblems.Itisangraphicalrepresentation ofallpossiblesolutionundergivenconditions.Itisalmost

similar to a tree because it has root node and corresponding branches. Cart algorithm is used to build tree. It is a popular Machine Learning method because it is easy to understandandinterpret,andtheycanhandlemultipleinputsandou tputs. They are often used for classification problems, but they can also be used for regression problems.

In a Decision Tree, the algorithms look for the most significantvariablesandrelationsamongthevariablesthatgive rise to the final Decision Tree. It splits the Population into 2or more than 2 homogeneous sets. This process is repeated on each derived set, called sub-population until we get the pure set or leaf node.

*2) Logistic Regression:*Logistic Regression is a statistical approach used to analyze the relationship between a binary response variable and one or more predictor variables. This method is commonly used to predict the probability of an event occurring based on a set of independent variables. Itusesalogisticfunction,alsoknownasthesigmoidfunction,to modeltherelationshipbetweentheindependentanddependent variables. The sigmoid function produces a probability value between 0 and 1 for any input value.

Thelogisticregressionmodelitestimatestheprobability of the dependent variable taking on the value of 1 (success) based on the independent variables. The coefficients of the independent variables are estimated using a maximum likeli- hood method to find the values that maximize the likelihoodof observing the data given the model. This model can make predictions by setting a threshold probability value above whichtheresponsevariableisclassifiedas1andbelowwhich it is classified as 0.

3) Support Vector Machine: The Support Vector Machine(SVM) is a powerful algorithm that is widely used for supervisedlearningtaskssuchasclassificationandregression. One of the key features of SVM is that it is a non-parametric algorithm, which means it does not make any assumptions about the distribution of the data The basic idea behind SVM is to find a hyperplane that can effectively separate the data into different classes. For binary classification, the hyperplane is a line that separates the data points into two groups basedon their class labels. This algorithm finds the margin between the two classes, which is the distance between the decision boundary and the closest data points from each class.

While SVM is a binary classification algorithm, it can handle multi-class classification problems using techniques such as one-vs-one and one-vs-all. SVM is a margin-based algorithm, which means it aims to maximize the margin between the decision boundary and the closest data points from each class. The optimization problem involved in SVM can be solved using techniques such as gradient descent and quadratic programming.

D. SystemFramework

The figure depicted in the description represents the step-by-step process of constructing and implementing a model. The first stage of this process entails the selection of a dataset thatwillbeutilizedtotrainthemodel.Inthisdataset,asubset known as the training dataset is set aside for model training, while another subset, known as the testing dataset, is used to assess the accuracy and effectiveness of the model.

There are various machine learning algorithms, such as Support Vector Machine, Logistic Regression, and Decision Tree, among others, that can be used to train the model. The algorithm that is most appropriate for the particular model being constructed is selected, and then utilized to train the model with the training dataset.

Afterthemodelhasbeenconstructed,itmustbetested to determine its accuracy and effectiveness in predicting and analyzing the data. The accuracy and effectiveness of the model rely heavily on the dataset selected, the algorithm utilized, and the training dataset employed.

Finally, once the model has been constructed and tested, it canbedeployedforuseinpredictionandanalysis.Itisimportanttonotethattheprocessofbuildinganddeployingamodel comprises multiple steps, and each stage must be carefully executed to ensure the model's accuracy and effectiveness.

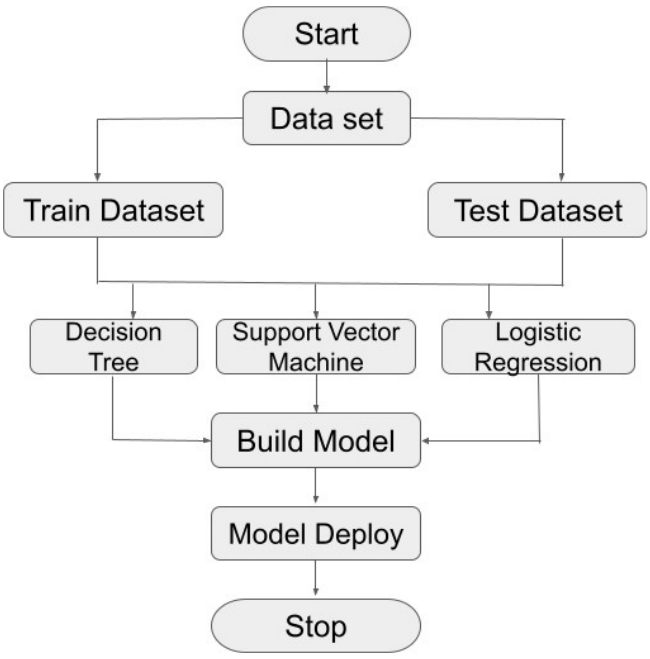

Fig.2.SystemFrameworkFlowchart

The project used a split of 80 percentage of the data for training and 20 percentage for testing. The three models, Logistics Regression, SVM, and Decision Tree, were trained using the training data. The performance of each model was then evaluated using the test data, and the Decision Treemodel was selected as the best-performing model based on accuracy. The selected model was then deployed using an appropriate production environment to ensure that it can be used by stakeholders to make accurate predictions. Deploying a model is an essential step in making it available for use by stakeholders,andtheprocessinvolvesseveralsteps,including

model selection, setting up a production environment, and monitoringthemodel'sperformanceovertime.Inthisproject, the Decision Tree was selected as the best-performing model basedonaccuracy.TheDecisionTreeisasupervisedmachine learning algorithm that is commonly used for classification and regression problems. It is a tree-like model where internal nodesrepresentfeaturesorattributes,branchesrepresentdecisionsorrules,andleavesrepresenttheoutcomeorprediction.

E. ModelDeployment

Modeldeploymentistheprocessofmakingatrainedmodel availableforuseinreal-worldscenarios.Inthiscaseof the project, the model deployment involves taking the best- performing model, which is the decision tree, and making it availableforusebystakeholders.Thisprocessinvolvesseveral steps, including saving the trained model in a suitable format, setting up a production environment to host the model, and exposingthemodelthroughanAPIoruserinterfacethat can be accessed by end-users. The deployed model must also be monitored to ensure that it is functioning correctly and providing accurate predictions.

In this project, the Decision Tree outperformed the other models, Logistic Regression and SVM, in terms of accuracy. This could be due to the Decision Tree's ability to handlenonlinear relationships and interactions between the features, which might have been present in the data. Moreover, it iseasytointerpretandvisualize,whichmakesthemuseful for understanding the decision-making process of the model. ThereforeDecisionTreealgorithmwasselectedbasedon its superior performance and its ability to handle non-linear relationships and interactions between the features.

IV. RESULT

A.confusionmatrix

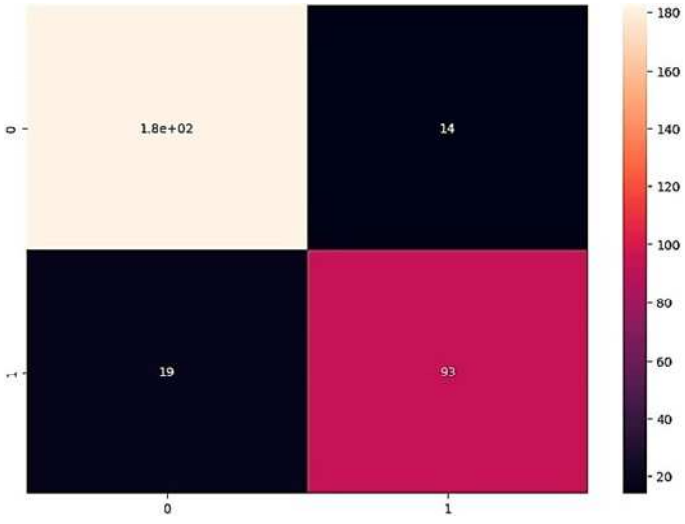

Fig.3.DecisionTree

The main objective of the dataset is to predict whether the patienthasdiabetesornot.Thedatasetismadebycombining

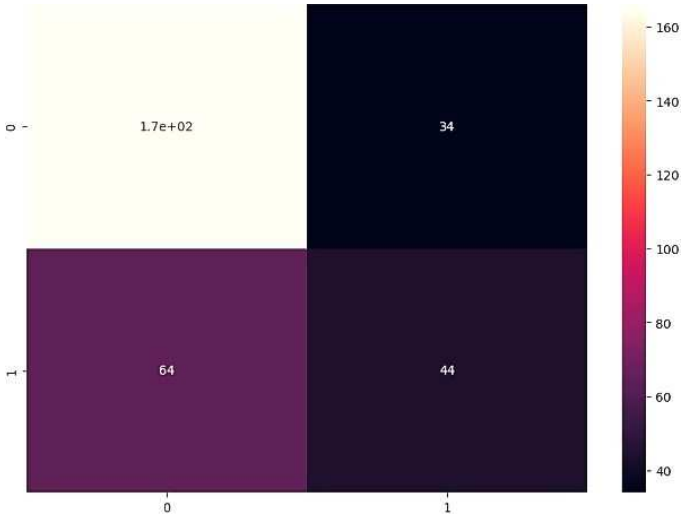

Fig.4.LogisticRegression

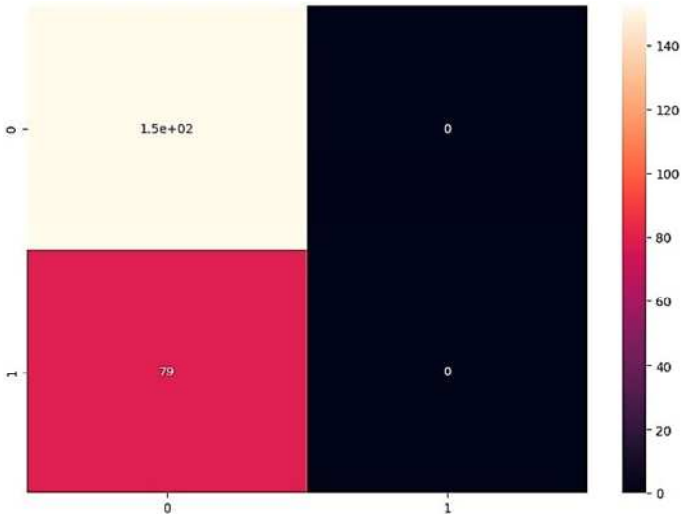

Fig.5.SupportVectorMachine

Pima Indian dataset (PID) and also the dataset which contains the food calorie consumption per day and heart rate which is made by us using the google form. And these are combined into one dataset. Based on this dataset diabetes prediction system is developed. The dataset is preprocessed and well cleaned in order provide great results. The program is made run on collab by using the above mentioned dataset. Here we are doing data visualisation in class distribution which will lookateachfeatureandfinallythecorrelationamongfeatures. The purpose of the testing is to check the errors and also to check which algorithm is more suitable to deploy the model. The algorithms used over here to test are Decision Tree, SupportVectorMachineandLogisticRegression.Thefeatures used to predict the best algorithms are accuracy, precision, recall and F1 score. By comparing the models we deploy the modelwhichprovidesgreaterresult.Thedeploymentofthe

modelwillbedonewiththeuseofDecisionTreealgorithm.

Model Name	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.88	0.86	0.83	0.84
Logistic Regression	0.68	0.56	0.40	0.47
Support Vector Machine	0.65	0.43	0.65	0.52

Table.2PerformanceMatrices

V. CONCLUSIONANDFUTURESOPE

Diabetes is a serious health problem that has an impact on social, psychological, and physical health. When the pancreas producesenoughinsulinbutthecellsdonotreacttoit,Type 2 diabetes develops. This increases blood sugar levels and raises the risk of developing various diseases, especially in women. In order to solve this problem, a system is developed toforecastaperson's(woman's)likelihoodofdevelopingType 2 diabetes. To assess performance and instal the best model, the system makes use of machine learning methods including Decision Trees, Support Vector Machines, and Logistic Regression [13]. The Pima Indian dataset (PID), heart rate and calorie consumption per day data were combined to create the dataset utilised to design this system. The application contains data visualisation and testing and is conducted on Google Colab.

Expanding the prediction models to include men and children requires using separate datasets tailored to their characteristics. Input features for each group may need adjusting to reflectthedifferencesinriskfactorsfordevelopingtype2diabetes.Personalizedassessmentscanimproveaccuracy.Clinical trials and experiments can improve the reliability and efficiency of predictive models for type 2 diabetes prediction. Close collaboration can refine and optimize the models to bettersuithealthcareprovidersandpatients,ultimatelyimproving diabetes prevention and management in the Malabar Area.

REFERENCES

[1] D. K. Malini M, Gopalakrishnan B, "Diabetic patient prediction usingmachine learning algorithm," IEEE, 2021.
[2] P. B. D. R. P. K.Harika, B.Ramya, "Diabetic prediction system usingmachine learning model," IEEE, 2022.
[3] Ashwini R, S M Aiesha Afshin, Kavya V, Prof. Deepthi Raj, "DiabetesPrediction Using Machine Learning"(IJRASET), Vol 10 issue 4, 2022.
[4] D. S. V. Mitushi Soni, "Diabetes prediction using machine learningtechniques," IJERT, 2020.
[5] F. K. Usama Ahmed, Muhammed Adnan Khan, "Prediction of diabetesempowered with fused machine learning," IEEE, 2021.
[6] M. R. AnnavarapuNaga Prathyusha, "Diabetic prediction using kernelbased support vector machine learning," ijatcse, 2020.
[7] J. R. A. Yahyaoui, A. Jamil and M. Yesiltepe., "A decision supportsystemfordiabetespredictionusingmachinelearninganddeeplearningtechniques," IEEE, 2019.
[8] D. V. v. Aishwarya mujumdar, "Diabetes prediction using machinelearning algorithms," 2019
[9] K. K. Ankur Goyal, "Analysis of various diabetic prediction methods ofmachine learning," ACM Digital Library, 2021.
[10] S. G. Neha Prerna Tigga, "Prediction of type 2 diabetes using machinelearning classification methods," 2020.
[11] Deepika Raj K.,Saani H., "Semi-automatic building of Domain Moduleby use of novel machine learning approach," 2015.
[12] Onur Sevli, "Diagnosis of diabetes using different classifiers," VOL38,2023.
[13] TriptiLamba,DrKavita,AKMishra,"OptimalMachinelearningModelfor Software Defect Prediction," 2019.
[14] S. Ramya, Dr. D. Kalaivani, "Machine Learning Approach for DiabetesPrediction," 2022.