JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 25 Issue 04, 2022 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

THE RULE-BASED MULTI-CLASS CLASSIFICATION MODEL PREDICTS EARLY DIABETES USING SUPERVISED MACHINE LEARNING TECHNIQUES

R. Karthikeyan[1a, b*], P. Geetha[2], E. Ramaraj[3]

[1a]Ph.D. Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, India-630003

[1b]Head i/c, Department of Computer Science, SRM Arts and Science College, Kattankulathur, Chennai, India-603203

[2]Associate Professor & Head, PG Department of Computer Science, Dr. Umayal Ramanathan College for Women, Karaikudi, India-630003

[3]Professor, Department of Computer Science, Alagappa University, Karaikudi, India-630003

*Corresponding Author Email: karthikeyan.r@srmasc.ac.in

Abstract

Diabetes is a metabolic disorder characterized by high blood sugar levels in which the body fails to create essential insulin or fails to utilize the insulin that is produced adequately. Diabetes can be caused by a failure to detect pre-diabetes early on. There were only two possible results in diabetes research previously: a Tested Negative or a Tested Positive result. The primary goal of this study is to identify pre-diabetes, as well as the Test Negative and Positive results, using a Rule-Based Multi-Class Classification Algorithm that can avoid the formation of Type II Diabetes. This research made use of the PIMA dataset. The variable relevance identifies the most important factors in the datasets such as BMI, Plasma glucose, and Blood pressure. The rules are developed based on the important variable. Using Supervised Machine Learning methods such as Decision Tree, RepTree and Logistic Regression approaches, the Rule-Based Multi-Class Classification model classifies and predict an individual them as Non-Diabetic, Pre-Diabetic, and Diabetic. Previous research has found limitations in Machine Learning Classifiers for Diabetes Prediction in terms of data size, accuracy, and multi-class predictor variables. The proposed system outperforms all of them and produces the best results in predicting Diabetes Mellitus in experimental data too.

Keywords: Supervised Machine Learning, Diabetes, Decision Tree, Logistic Regression, RepTree, Rule-Based Multiclass.

I. INTRODUCTION AND MOTIVATION

Prediabetes is an early symptom of diabetes. This study intends to detect the precision diabetic illness in advance by utilizing the Rule-Based System with Multi-Class Classification Model [RBSMCC], which is used in this research work. Pre-diabetes can affect anybody who lives an unhealthy lifestyle. If pre-diabetes is discovered early by lab tests results, it can be cured by modifying one's lifestyle patterns. Previous studies on diabetes prediction have only used binary groups as tested positive and tested negative categories.

Submitted: 21/10/2022 **Accepted**: 24/11/2022

A person who tests positive for diabetes may be in the pre-diabetic stage of the disease, known as Pre-diabetes. Diabetes tests, such as the Oral Glucose Tolerance Test and Fasting Plasma Glucose, can assist in confirming this. Blood sugar levels are classified into three categories: (i) HIGH i.e. above 6.5 percent (above 48 mmol/mol approx.) it may be a risk for diabetes, (ii) MODERATE i.e. 5.7 percent - 6.4 percent (39 to 46 mmol/mol approx.) it may be a risk for pre-diabetes, or (iii) NORMAL i.e. less than 5.7 percent (39 mmol/mol approx.) The majority of prior diabetes research outcomes Asha Gowda Karegowda et al. [**25**], A. Mary Posonia et al. [**26**], and Rashedur M. Rahman et al. [**28**] were based on plasma variables as a root node in decision trees. When compared to plasma, the BMI variable has a high significance, as evidenced by using the Feature selection using Information Gain Ranking Filter.
A Body Mass Index (BMI) more than 24 indicates an unhealthy physique and may indicate pre-diabetes. This can be performed using Supervised Machine Learning methods like Logistic Regression, Decision Tree J48, and RepTree classifiers. An RBSMCC is a data-driven strategy in which diabetes mellitus facts are gathered using a knowledge base, i.e., diabetes standards from the World Health Organization, and rules are applied to important data to achieve the purpose of creating facts from it. People with pre-diabetes can be easily cured if their blood sugar levels are moderate, i.e., FPG 100-125 mg/dl. People with pre-diabetes should lose weight; eat a variety of nutrients, and exercise regularly to stay fit.

II. BACKGROUND AND RELATED WORK

Diabetes may be diagnosed using a variety of machine learning techniques, as explained in the studies below. Leon Kopitar et al. focused on the variable importance for five models such as Linear Regression, Random Forests, eXtreme Gradient Boosting (XGBoost), regularized generalized linear model (Glmnet), LightGBM are compared with predictor variables to get better performance using fasting plasma glucose level [24]. Mahboob Alama et al. used three methods for diabetes prediction: K-means Clustering, Artificial Neural Network and Random Forest. In addition, the report reveals that the major variables BMI and blood glucose are more important for diabetes prediction [**7**]. Su Su Maw et al. examined the influence of Glycated Haemoglobin levels in middle-aged and aged Japanese adults' diet between supper and night using anthropometric and lifestyle data from people aged 40–74 years with no diabetes symptoms [**2**]. Vandana Rawat et al. examined five machine learning approaches, namely Bagging, LogicBoost, RobustBoost, AdaBoost, and Naive Bayes, using the PIMA dataset for diabetes mellitus prediction, and Bagging performed well when compared to other machine learning algorithms [**11**]. Prema NS et al. examined 10 classifiers, including KNN, Logistic Regression, Decision Tree, Naive Bayes, Linear SVM, RBF SVM, Gaussian Process, AdaBoost, Random Forest, and Voting Classifier 30 percent test data, to predict diabetes using ensemble approaches on the PIMA dataset [**17**]. Komal Patil et al. focused on a hybrid model comprising a Decision Tree, XGBoost, and Voting Classifiers to diagnose diabetic illness and evaluated other machine learning methods [**23**]. Zou Q et al. [**18**] employed three classifiers, including a Random Forest, a Decision Tree J48, and a Neural Network, to predict Diabetes Mellitus using two datasets, PIMA and Luzhou. They further validated the model with 5-fold

cross validation and compare crucial aspects for diabetes prediction; Principal Component Analysis (PCA) was performed. Talha Aakansha Rathore et al. [**9**] the concept is to detect and predict diabetes disorders in women with PIMA dataset with Support Vector Machine and Decision Tree, machine learning techniques. The R framework has been utilized for diabetes prediction. Shengqi Yang et al.2017[1] employed data mining methodologies to predict type II diabetes using the Logistic Regression algorithm, K-means using the diabetes dataset to enhance accuracy, and it also works on numerous datasets to evaluate the model's performance. D. Ashok Kumar et al. [**21**], compared five classification techniques such as BayesNet, Decision Table, Naive Bayes, Regression, and SVM to classify the diabetes with feature selection and all attributes as a hybrid model to get better accuracy. Beata Strack et al. [**19**] used Multivariable Logistic Regression to evaluate HbA1c and other risk variables for diabetes mellitus readmission. The statistical model predicts the same findings. One of the most important aspects in diabetes management is the HbA1c test. Md. Aminul Islam et al. [**29**] focuses on Machine Learning classifiers such as Naive Bayes, Logistic Regression, Multilayer perception, Support Vector Machine, K-Nearest Neighbor, AdaBoostM1, Bagging, OneR, J48, Random Forest are used to predict the onset of diabetes, in which the logistic regression scored 78.01 percent accuracy as best among ten classifiers.

The majority of the study focused on positive and negative results in diabetes data, and there was a scarcity of research on pre-diabetes, which is highly essential to detect diabetes and may help diabetic individuals prevent problematic diabetes.

III. PROPOSED METHODOLOGY

This section included the dataset, the model, feature selection based variable importance, Diabetic Screening Guidelines & Rules, rule establishing, and supervised machine learning approaches with multi-Class classification.

A. DATASET DESCRIPTION

The dataset was obtained from the PIMA Indians diabetes data repository at the University of California, Irvine. Pregnancies, Glucose, Age, BMI, Diastolic Blood Pressure, Diabetes pedigree function, Insulin, Triceps Skin Fold Thickness are the eight independent variables and one outcome variable binary class in the PIMA dataset; only Diabetes pedigree function and Age do not have any missing data.

B. ARCHITECTURE DIAGRAM

The PIMA dataset of 768 training data is tested using a Binary class with three machine learning classifiers, Logistic Regression, Decision Tree and RepTree and the models are validated using a confusion matrix. If the model precision falls below 95%, retest using Feature selection using Information Gain Ranking Filter and apply rules to improve the model with multi-class classification training data using the three classifiers as shown in Figure 1. When a model achieves a high degree of accuracy through performance metrics it is referred to as a predictive model. Then it is evaluated using new test data or previously unknown data to

validate the model. The Rule-Based Multi-Class classification approach assesses whether the suspect has Prediabetes, Diabetes, or is otherwise healthy.

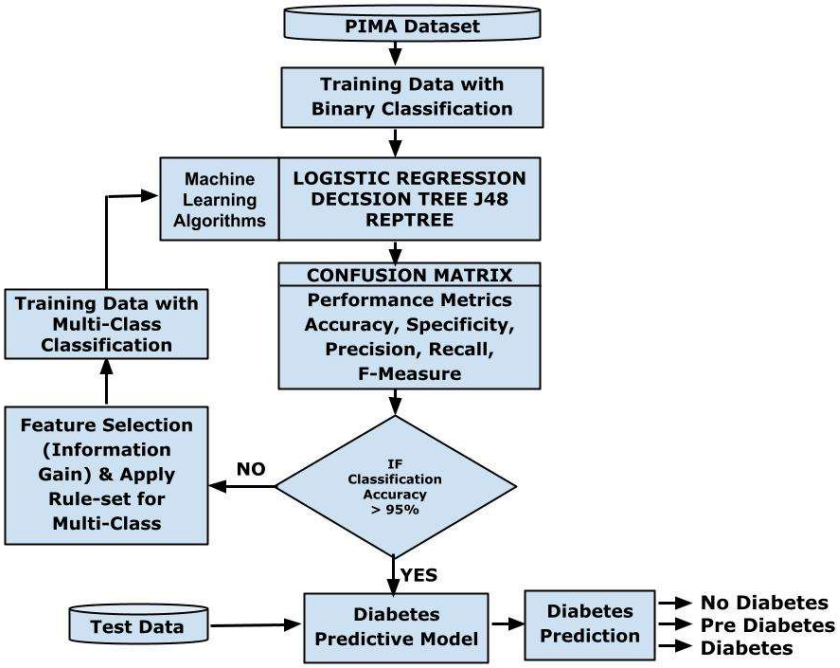

Figure 1: Multi-Class Classification Model

C. FEATURE SELECTION USING INFORMATION GAIN RANKING FILTER

The comparisons of Binary-Class versus Multi-Class utilizing Feature Selection (Information Gain) obtained by the attribute evaluator using the WEKA tool are shown in Figure 2. The Multi-Class classification has a greater chance of predicting diabetes disease than the Binary-class based on five essential variables: BMI, Blood Glucose, Blood Pressure, Skin Thickness, and Insulin.

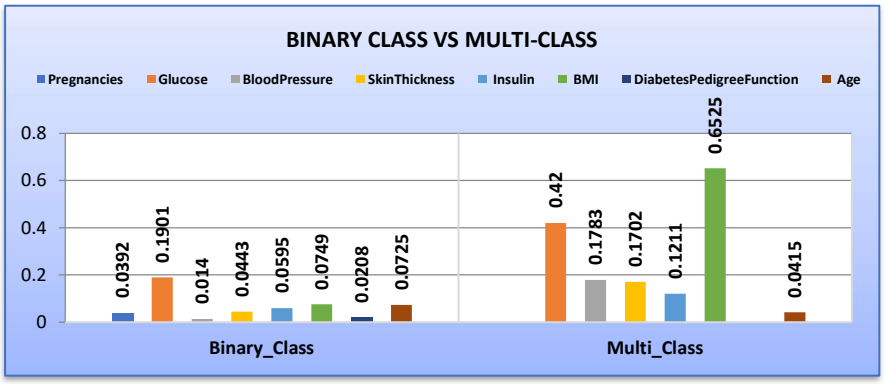

Figure 2: Binary-Class vs. Multi-Class Comparisons

Figure 2 illustrates the variable significance for the PIMA dataset for binary-class and Rule-

Based Multi-Class utilizing Information Gain Ranking Filter. The variable importance has a significant influence in obtaining good outcomes for classification-based problems Leon Kopitar et al. [**24**]. When the best ranking variable is picked as a root variable, the accuracy is likewise high. When the key variable in the data set is used to forecast illnesses such as diabetes, the likelihood of obtaining high accuracy increases.

D. DIABETIC SCREENING GUIDELINES

Diabetes does not occur abruptly in the human body; prior to diabetes, they are frequently afflicted by pre-diabetes; if not correctly detected at an earlier stage, they will progress to chronic diabetes. Diabetes medical assessment may be carried out in three aspects: medical history, physical examination, and laboratory evaluation, which includes seven tests such as FPG, OGTT, HbA1c, lipid test, liver test, urine test, and thyroid test specifically for women, American Diabetes Association [**8**]. The majority of people with pre-diabetes are asymptomatically or unwittingly affected; it may only be detected with a lab test, or others may have acquired symptoms, although it is not frequent in all people. Blood glucose (sugar) levels are greater than normal in pre-diabetes, although the condition is not diagnosed as diabetes. Diabetes can be detected more quickly if the fasting plasma glucose or 2-hour plasma glucose value is combined with a 75g oral glucose tolerance test or the A1C test [**3**].

The nomenclature of the Diabetes Screening Standards categories are as follows: normal, pre-diabetes, and diabetes.

FPG **<=**99,100 To 125, >=126.
ORAL GLUCOSE **<=**139, 140 To 199, >=200.
PPBG **<=**139, 140 To 199, >=200.
DIASTOLIC BP **<=**80, 80 To 90, >=91.
SYSTOLIC BP **<=**120, 120 To 139, >=140.
BMI **<=**24, 25 To 29, >=30.
HBA1C **<=5**.6, 5.7 To 6.4, >=6.5.
RANDOM BLOOD SUGAR **>=**200.

1. ESTABLISHING RULES

A rule extraction is developed based on the diabetes screening standards presented above. As part of the rule-making process, eight tests for diabetes screening are examined. The first six rules (A to F) recommend the value for normal range for diabetes screening, Rules (G to L) represent the pre-diabetes range, and Rules (M to S) reflect the diabetes range. YoichiHayashi et al. [**10**] proposed a novel white box model rather than a black box model for diabetes prediction using the PIMA dataset. The white box model combines Recursive-Rule extraction (Re-RX) and J48graft to give accurate diabetes categorization. Anand Kumar et al. created a diabetes monitoring system consisting of a set of rules and a patient's health data as an app that is deployed online, and the results were compared with other approaches for diabetes prediction such as Neural Network, Logistic Regression, SVM, RBF-SVM, and Decision Tree in Anand Kumar Srivastava et al.[**12**]. Najmeh Hosseinpour et al. [**16**], outlined five classifiers for

diagnosing diabetes using the PIMA dataset: Ensemble, Rule-base, Bayesian, Functional, and Decision Trees.

2. DIABETES MELLITUS IDENTIFICATION CRITERIA

The types of diabetes diagnoses are listed below with names such as normal, pre-diabetes, and diabetes**.**

NORMAL

Rule A: Fasting Plasma Glucose (FPG) between 70 and 99 mg/dL (3.9 to 5.5 mmol/L),
Rule B: Postprandial blood glucose (PPBG) less than 140 mg/dL (7.8 mmol/L),
Rule C: HbA1c less than 5.7 percent (39 mmol/L approx.),
Rule D: Systolic blood pressure (SBP) less than 80,
Rule E: Diastolic blood pressure (DBP) less than 120.
Rule F: Body mass index (BMI) less than 24.

PRE-DIABETES

Rule G: Fasting Plasma Glucose between 99 and 126 mg/dL (5.5mmol/L to 7.0 mmol/L).
Rule H: PPBG or OGTT between 140 and 200 mg/dL (7.8 and 11.1 mmol/L).
Rule I: HbA1c between 5.7 and 6.4 percent (39 to 46 mmol/L approx.).
Rule J: Systolic blood pressure between (SBP) 120 and 139.
Rule K: Diastolic blood pressure (DBP) between 80 and 89.
Rule L: Body mass index (BMI) between 25 and 29.

DIABETES

Rule M: Fasting Plasma Glucose levels more than 126 mg/dL (7.0mmol/L)
Rule N: PPBG or OGTT of 200 mg/dL (11.1 mmol/L) or higher.
Rule O: HbA1c greater than 6.5 percent (about 48 mmol/L).
Rule P: A plasma glucose level of 200 mg/dL (11.1 mmol/L) or greater at random.
Rule Q: SBP greater than 139.
Rule R: DBP greater than 89.
Rule S: BMI greater than 29.

3. EXTRACT RULES FOR DIABETES MELLITUS IDENTIFICATION

Symptomatic threshold values are recommended as cut-off points for the diagnosis of undiagnosed diabetes, prediabetes, and diabetes. The Diabetes Mellitus identification criteria nineteen rules are combined to form the three rules for diagnosing diabetes. When the FPG, PPBG, DBP, BMI, or HbA1c Rule 1 test criteria match, the individual is considered healthy. If the FPG, PPBG, or OGTT, DBP, BMI, or HbA1c Rule 2 test criteria match, the individual has pre-diabetes. When the Rule 3 FPG, PPBG, or OGTT, DBP, BMI, or HbA1c test requirements are met, diabetes is diagnosed.

RULE 1: If ((FPG <=99) and (PPBG <=139) and (DBP<=80) and (BMI<=24) or (HbA1c<=5.6)) → NORMAL

RULE 2: If ((FPG range from 100 to 125) and (PPBG or OGTT range from 140 to 199) and (DBP range from 80 to 90) and (BMI range from 25 to 29) or (HbA1c range from 5.7 to 6.4))

→ PRE-DIABETES
RULE 3: If ((FPG >=125) and (PPBG or OGTT>=200) and (DBP>=90) and (BMI>=30) or (RBS>200) or (HbA1c >=6.5)) → DIABETES

IV. METHODOLOGY - MACHINE LEARNING ALGORITHMS

This section explains the Decision Tree, Logistic Regression, RepTree supervised machine learning classifiers, which are used to predict diabetes mellitus using a Rule-Based system with a multi-Class response variable to determine how many women have pre-diabetes, diabetes, or no diabetes.

4.1.1 DECISION TREE

The Decision Tree has the advantage of being a top-down strategy that supports non-linearity and has improved accuracy when used with classification and regression issues. Impurities such as Entropy and Gini can be used as criteria metrics to validate the Decision Tree. It is a discriminative model, and with the supplied data, accurate decision tree might be developed, a more flexible. The Decision Tree is most typically used when the target or response variable is categorical; it applies rules to form the tree for the independent or predictor variables, with the target variable functioning as a class variable. The Decision Tree's root node will be the most important variable, with following nodes constructed based on a condition over a feature. A Decision Tree can quickly identify a medical issue. Gaganjot Kaur et al. [**20**] focused on the Decision Tree J48 algorithm to identify and forecast diabetes using the PIMA dataset with a 99.87 percent accuracy. R. Sengamuthu et al. [**22**] conducted a comparative analysis using fourteen articles linked to the PIMA dataset using several classifiers, and found that Decision Tree J48 had the highest accuracy when compared to other classifiers. Dongmei Pei et al. [**13**] used a Decision Tree J48 algorithm for a diabetes-prediction model based on fourteen risk variables connected with diabetes. The Decision Tree is constructed using twenty rules. BMI is the most significant independent variable among the fourteen selected as the root node.

4.1.2 LOGISTIC REGRESSION

In general, Logistic Regression algorithms handle binary classification issues and estimate discrete variables' probabilities as 0/1, True/False, or Yes/No. It also enables multi-Class classification, which is referred to as multinomial regression. The logistic regression algorithm operates on the basis of a cutoff value, i.e., if X is 0.5, the result is 0 otherwise it is 1. Priya. M et al. [**15**] examined three classification methods, such as Decision Tree, Naive Bayes, and Logistic Regression, for diabetic prediction at an initial stage using diabetes dataset of 372 occurrences, with Logistic Regression outperforming the other two. Changsheng Zhua et al. proposed a diabetes prediction model based on data mining using the diabetes dataset, which includes principal component analysis to reduce dimensionality, to locate outliers and eliminate wrongly input data, used K-means, and Logistic Regression to integrate the model for improved accuracy [**6**].

4.1.3 REDUCED ERROR PRUNING TREE (REPTREE)

Reduced Error Pruning (REP) is also called as Rep Tree, it is an extension of c4.5 method to improve the pruning phase and generate a decision tree. It is also considered as fast decision tree learner based on information gain. Gaganjot Kaur et. al. in [20], presented with decision tree rep tree, J48 algorithm to classify and predict diabetes using PIMA dataset with better accuracy. Results suggest that Insulin also has importance in diabetes prediction

4.1 ALGORITHM FOR DIABETES PREDICTION USING MULTI-CLASS CLASSIFICATION

Input: Training dataset (T) of PIMA Indians diabetes data with eight attributes.
Output: (i) Binary-Class (BC) Decision Tree with two values Tested Positive, Tested Negative,
(ii)Multi-Class (MC) Decision Tree with three values such as Normal, Pre-diabetes, and Diabetes.
Learning Model = Logistic Regression (), Decision Tree J48 (), RepTree().
Eval Model = Training Data.
Step-1: Preprocess the dataset (T) with BC.
Step-2: Choose classifier and score the model
(i) For (M=0; M≤2; M++) do
(ii) Model = Learning Model (M);
(iii) Eval Model by Training Data;
(iv) Performance Analysis by Confusion Matrix Metrics (Accuracy (M));
Step-3: If (Classification Accuracy > 95 %) Then go to Step-5
Else Step-4;
Step-4: Rule-Based Multi-Class Classification model creation:
(i) Apply variable Importance for Training Data (T) using Information Gain to find the three best parameters such as BMI, Plasma Glucose, and Blood Pressure;
(ii) Extract Rules for the three key parameters to find multi classes such as Normal, Pre-Diabetic, Diabetes (MC);
(iii)Apply MC with training dataset T to get a multi-Class response variable.
(iv)**Go to** Step-2;
Step-5: Test new data without outcome variable to validate the model which predicts the suspected person is affected by prediabetes, diabetes or not;

4.2 MULTI-CLASS CLASSIFICATION

Multi-Class Classification is described as having more than two classes; in this study, three classes are used to forecast diabetic mellitus diseases such as Diabetes, Pre-diabetes, and No diabetes from diabetes data by extracting appropriate IF-THEN rules to classify records. The three integrated rules and two classifiers are evaluated to predict whether asymptotic people have Pre-diabetes or Diabetes by using a diabetes dataset. The target or response variable is separated into three types: No diabetes, Pre-diabetes, and Diabetes, and it is represented by a

3x3 confusion matrix. The computation of FP: False Positive, FN: False Negative, TP: True Positive, TN: True Negative in multi-class classification is not as fixed as it is in binary classification. Each class has its own 3x3 confusion matrix, computation table 1 demonstrates this. The TP value in class 1 is positioned at the top left, class 2 in the center, and class 3 in the bottom right, with all other TN, FP, and FN values located in various locations. A.Tharwat et al. [**14**] introduced classification assessment metrics such as binary classification, multi-class classification utilizing scalar values, and graphical measures such as accuracy, sensitivity, specificity, and ROC, and precision, recall for a better interpretation of any model. The above six accuracy measurements were used by Deepti Sisodiaa et al. [**4**].

Table 1 | Computation of the confusion matrix for Class1, Class 2, Class 3 (TP, TN, FP, FN).

		ACTUAL VALUE		
PREDICTED VALUE		Class 1	Class 2	Class 3
	Class 1	TP	FP	FP
	Class 2	FN	TN	TN
	Class 3	FN	TN	TN

		ACTUAL VALUE		
PREDICTED VALUE		Class 1	Class 2	Class 3
	Class 1	TN	FN	TN
	Class 2	FP	TP	FP
	Class 3	TN	FN	TN

		ACTUAL VALUE		
PREDICTED VALUE		Class 1	Class 2	Class 3
	Class 1	TN	TN	FN
	Class 2	TN	TN	FN
	Class 3	FP	FP	TP

The Confusion Matrix is used to assess a classifier's performance and to determine how accurate a classifier is producing classification predictions. The confusion matrix for binary and multi-class classification for diabetes prediction using two classifiers is shown below in Table 2, along with training and k-fold cross validation data. Among the two classifiers, the Decision Tree J48 performed well, and accuracy improved.

Table 2 | Diabetes Confusion Matrix for Binary vs. Multi-class Classification

	BINARY CLASSICATION			MULTI-CLASS CLASSIFICATION			
DECISION TREES (J48) - Training Data	Actual Class			Actual Class			
	Correct	646	84.11%	Correct		761	99.08%
	Incorrect	122	15.89%	Incorrect		7	0.92%
	===Confusion Matrix ===			a	b	c	<-- Classified as
Predicted Class	a	b	<-- Classified as	175	0	0	a = DIABETES
	178	90	a = tested positive	1	228	5	b = NO DIABETES
	32	468	b = tested negative	1	0	358	c = PRE-DIABETES

LOGISTIC REGRESSION Training Data	Correct	601	78.26%	Correct		656	85.41%
	Incorrect	167	21.74%	Incorrect		112	14.59%
				a	b	c	<-- Classified as
	a	b	<-- Classified as	140	1	34	a = DIABETES
Predicted Class	156	112	a = tested positive	1	204	29	b = NODIABETES
	55	445	b = tested negative	22	25	312	c = PREDIABETES

REPTREE Training Data	Correct	641	83.46%	Correct		760	98.95%
	Incorrect	127	16.53%	Incorrect		8	1.05%
				a	b	c	<-- Classified as
	a	b	<-- Classified as	138	1	36	a = DIABETES
Predicted Class	153	115	a = tested positive	1	203	30	b = NO DIABETES
	60	440	b = tested negative	25	28	306	c = PRE-DIABETES

V. RESULTS AND DISCUSSION

The Rule-Based Multi-Class classification Decision Tree is generated with better accuracy by employing BMI as a root node, which is one of the critical factors for diabetes prediction

illustrated in Figure 3. The Decision Tree divides the dataset in several ways by applying conditions and rules to the variables BMI, glucose, and blood pressure. Figure 3 displays the output of a WEKA Decision Tree for diabetes prediction built for multi-class classification, using BMI as the root node at depth zero, plasma glucose variable at depth one, and blood pressure variable at depth two to create three unique leaves. NODIABETES, PREDIABETES, and DIABETES

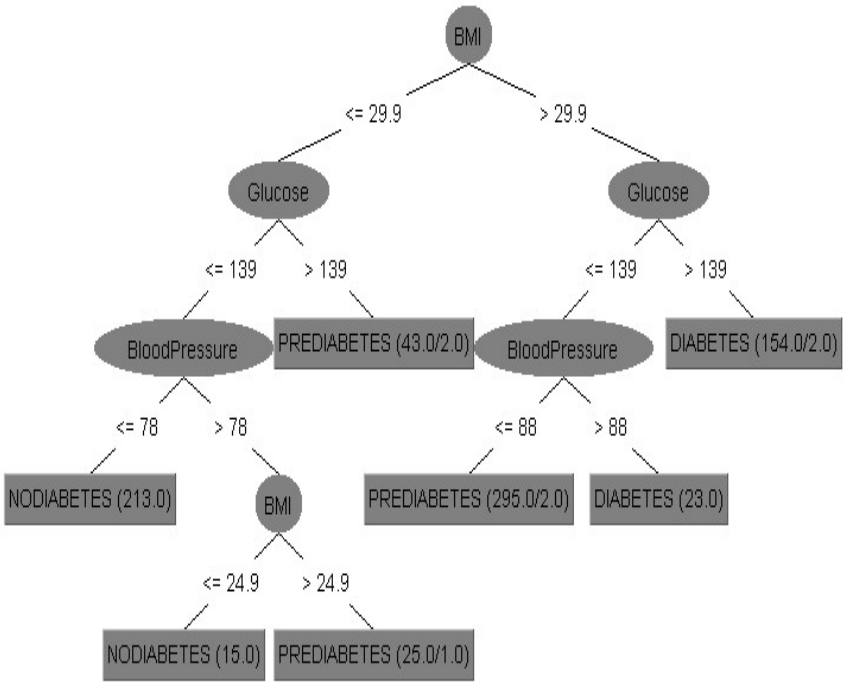

Figure 3: Diabetes Prediction Using a Decision TreeJ48 with Multi-Class Classification

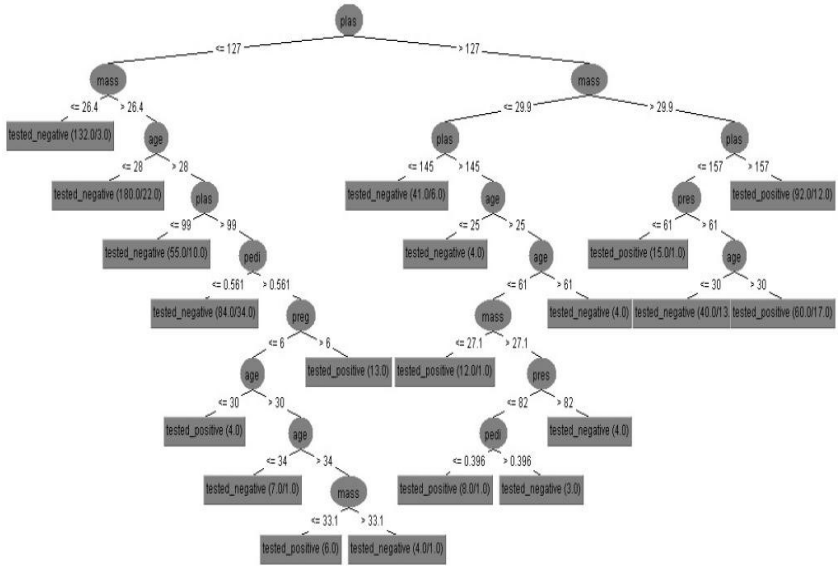

Figure 4: Diabetes Prediction Using a Decision TreeJ48 Based on Binary-Class Classification

Figure 4 displays the Decision Tree generated by binary classification, which is a difficult to grasp node separation. Figure 4 shows a more complicated Decision Tree created using binary classification applying the plasma variable as a root node. In binary class Glucose, BMI, and age are important, and Decision Trees are built based on glucose variables, but in multi-class; BMI, glucose and blood pressure are important variables, and Decision Trees are built based on BMI variable.

Table 3 |Compares three classifiers employing binary classification with multiclass classification training data.

Classifiers	BINARY CLASSIFICATION	MULTI-CLASS CLASSIFICATION
Decision Tree (J48)	84.11%	99.09%
RepTree	83.46%	98.95%
Logistic Regression	78.26%	85.42%

Table 3 compares Binary classification and multi-class classification for diabetes

Prediction using three supervised machine learning techniques: Decision Trees J48, Logistic Regression, and RepTree. Among the three classifiers, the Decision Tree J48 classifier achieved a high accuracy of 99.09 percent, RepTree scored 98.95 percent in multi-class classification, and Binary classification scored 84.11 percent, 83.46 percent, and 78.26 percent in predicting diabetes with tested positive and tested negative.

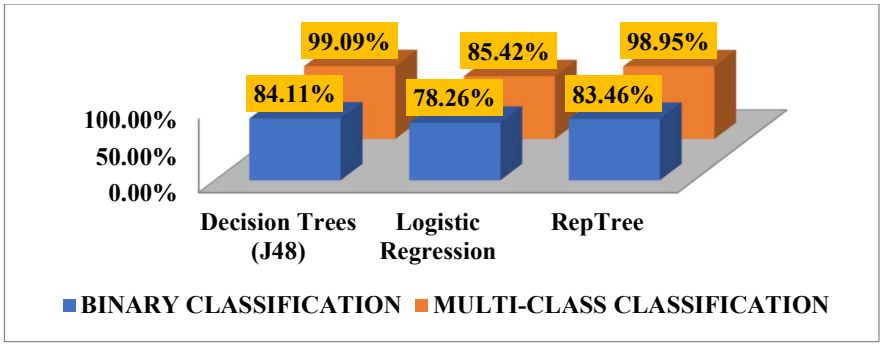

Figure 5: Training data comparison of three classifiers with binary and multi-class classifications

Figure 5 compares the binary class versus multi-class accuracy percentage of training data for

diabetes prediction among three classifiers, with Decision Tree J48 having the best accuracy percent 99.09. Accuracy measurements like (TP) True Positive rate, (FP) False Positive rate, (P) Precision, (R) Recall, F-Measure, and ROC were employed to verify the models. Five different metrics for measuring model validity can be calculated from the confusion matrix.

ACCURACY = (TN + TP) / (TN + TP + FN + FP)
SPECIFICITY = TN / (TN + FP)
PRECISION = TP / (FP +TP)
SENSITIVITY (or) RECALL = TP / (TP + FN)
F1-SCORE = F1 = 2TP / (2TP + FP + FN)

Table 4 | Detailed Accuracy of Decision Tree (J48) by Class for Diabetes Disease Prediction

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0.003	0.989	1	0.994	0.993	0.999	0.99	DIABETES
0.974	0	1	0.974	0.987	0.982	0.992	0.99	NODIABETES
0.997	0.012	0.986	0.997	0.992	0.984	0.996	0.991	PREDIABETES

Table 5 | Comparison of Weighted average of Three Classifiers in Multi class Classification

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
J48	0.991	0.006	0.991	0.991	0.991	0.985	0.995	0.99
Logistic Regression	0.854	0.096	0.855	0.854	0.854	0.764	0.952	0.92
RepTree	0.99	0.006	0.99	0.99	0.99	0.983	0.994	0.99

Table 6 | Multi class versus Binary class

	Tested Positive	Tested Negative	Multi-Class Classification Total
DIABETES	125	50	175
NODIABETES	27	207	234

PREDIABETES	**116**	**243**	**359**
Binary Classification Total	**268**	**500**	**768**

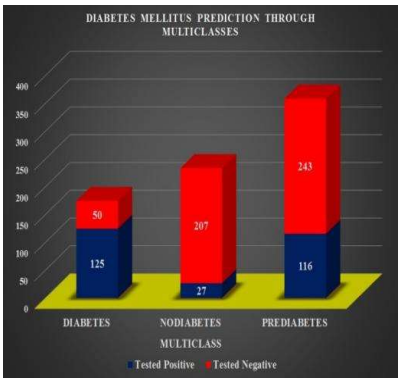

Figure 6: Prediction of Multi-class with Binary Class Occupancy

According to the results of rule-based multi-class classification predictions for diabetes shown in table 6, 175 women were categorized as diabetic, 359 as pre-diabetic, and 234 as normal, i.e., no diabetes. Figure 7 depicts the comparisons of the existing system and the proposed system. It demonstrates that Decision Trees outperformed Logistic Regression in multi-class prediction of diabetes mellitus.

Table 7 | Advantage of Proposed system with Multiclass classification

S.NO	ALGORITHMS	REFERENCE	DATA SET SIZE	ACCURACY	CLASSIFICATION TYPE
1	**Decision Tree J48**	**Proposed Method**	**768**	**99.08%**	**Multi-Class**
2	**RepTree**	**Proposed Method**	**768**	**98.95%**	**Multi-Class**
3	**Logistic Regression**	**Proposed Method**	**768**	**85.41%**	**Multi-Class**

The proposed system with Multiclass Classification is shown in Table 7, where the Decision Tree scored 99.08 percent, RepTree scored 98.95 percent, and Logistic Regression scored 85.41 percent accuracy. The proposed system outperforms the existing system with binary classification, as shown in table 8, in which two or three classifiers are combined to achieve the highest accuracy of 97.40 percent, whereas multiclass classification achieved 99.08 percent using a single classifier Decision Tree J48 with 768 data from the PIMA diabetes dataset.

Table 8 | Comparisons of Existing system with Binary Classification

S.NO	ALGORITHMS	REFERENCE	DATA SET SIZE	ACCURACY	CLASSIFICATION TYPE
1	PCA + K-Means + Logistic Regression	Changsheng Zhua et al. (2019)[6]	768	97.40%	Binary
2	K-Means cluster + Logistic Regression	Han Wu et al.(2018)[1]	589	95.42%	Binary
3	J48 Graft	YoichiHayashi n et al. (2016)[10]	768	84.97%	Binary
4	Random Forest	Neha Prerna Tiggaa(2019) [5]	768	75.00%	Binary
5	Decision Tree J48	Deepti Sisodiaa et al.(2018) [4]	768	73.82%	Binary

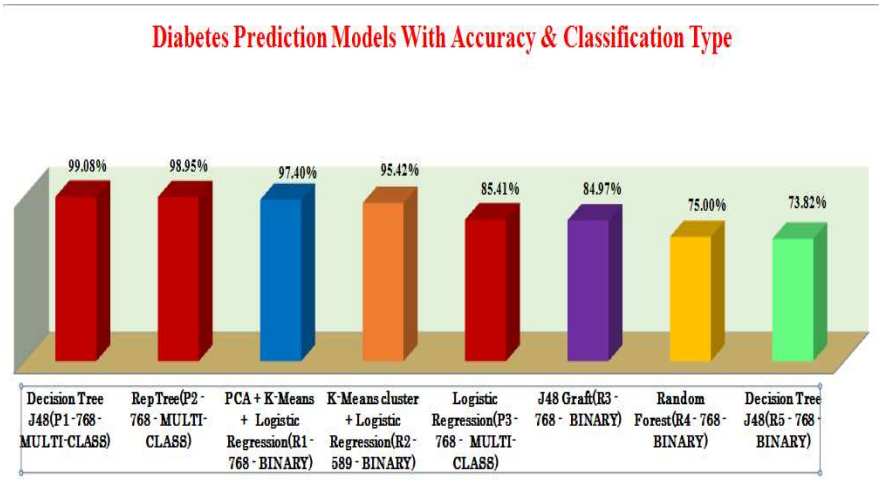

Figure 7 Comparisons of Proposed System (P1, P2, and P3) vs. Existing System (R1 to R5)

Figure 7 shows multiclass classification in red and binary classification in different colors. The decision tree J48 outperformed all previous works in diabetes prediction, with an accuracy of 99.08 percent. The figure above shows that the multiclass classifier works better than the binary

classifier in predicting diabetes for the suspect. It may be the best model for early detection of diabetes and prevention of the development of diabetes. Diabetes may be detected early in individuals by laboratory test results, which is an important step in diabetes prevention.

Table 8 | Test data or unseen data for the prognosis of diabetes without results

S. No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Multi Class
1	2	142	82	18	64	24.7	0.761	21	NA
2	6	144	72	27	228	33.9	0.255	40	NA
3	1	71	48	18	76	20.4	0.323	22	NA
4	6	93	50	30	64	28.7	0.356	23	NA
5	1	122	90	51	220	49.7	0.325	31	NA

The Table 8 exhibits the test data for diabetes prediction without class variables (NA - Not Available). The test data is fed into three different supervised machine classifiers: Logistic regression, Decision Tree J48, and RepTree. Table 9 displays the test data results. .

Logistic regression **Decision Tree J48** **RepTree**

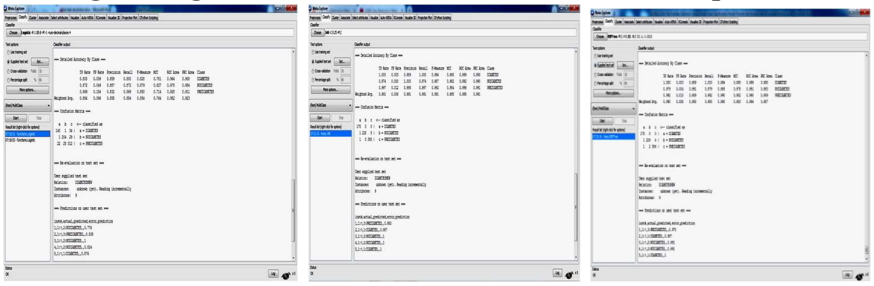

Figure 8 | Test Set Result of Logistic Regression, Decision Tree J48, RepTree

Table 9 | Test Data Results

S.NO	LOGISTIC REGRESSION	DECISION TREE J48	REP TREE
1	NODIABETES	PREDIABETES	PREDIABETES
2	PREDIABETES	DIABETES	DIABETES

3	**NODIABETES**	**NODIABETES**	**NODIABETES**
4	**NODIABETES**	**NODIABETES**	**NODIABETES**
5	**DIABETES**	**DIABETES**	**DIABETES**

Table 9 demonstrates that Decision Tree J48 and RepTree give similar results, while the Logistic Regression differs in the first two rows. Two of the three supervised machine learning classifiers indicate that the five unseen experimental data sets are identical.

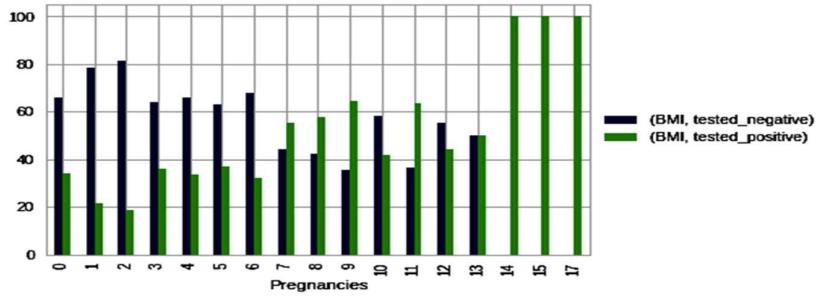

Figure 9: BMI and Pregnancy with Binary Classification

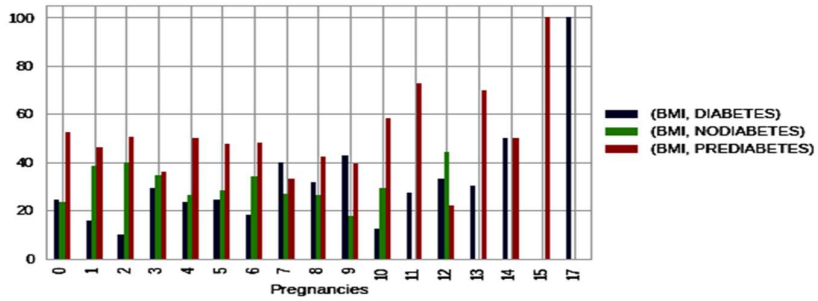

Figure 10: BMI and Pregnancy with Multi- Class Classification

Figure 9 and 10 depicts a comparison of binary classification versus multi-class classification using the PIMA dataset with two variables, BMI and pregnancy. As shown in Figure 10, the proposed multi-class classification creates a decision tree by BMI, which identifies women with pre-diabetes and takes adequate precautionary measures to prevent the development of diabetes. The Binary classification, as shown in Figure 9, is flawed in predicting prediabetes. PIMA Women who were 0 to 17 times pregnant were more likely to acquire prediabetes, as shown red in Figure 10. People with early stage of diabetes can prevent developing chronic diabetes if they are identified and treated early.

VI. CONCLUSION

Early pre-diabetes prediction is accomplished using Rule-Based Multi-Class classification utilizing three supervised machine learning methods, such as Logistic Regression, RepTree and Decision TreeJ48 in which pre-diabetes, diabetes, or no-diabetes are detected in patients using

PIMA dataset and experimental data. Previously, the binary classifier produced either positive or negative results. Most prior diabetes research lacks pre-diabetes-based predictions; since the Rule-Based Multi-Class Classification model is built and tested using three classifiers. The Decision Tree J48 classifier functioned admirably, and its accuracy has significantly increased to 99 percent. Finding prediabetes early in asymptotic or probable patients with a diabetes test has been regarded as a critical objective for preventing diabetes mellitus worldwide.

Declarations
Funding : Not applicable.
Informed Consent Statement : Not applicable.
Data Availability Statement: Not applicable.
Conflicts of Interest : The author declares no conflict of interest.

Acknowledgment
This article is submitted to this journal, and it has not been submitted to any of the journals and conferences anywhere before.

REFERENCES

[1] Han Wu, Shengqi Yang *, Zhangqin Huang, Jian He, Xiaoyi Wang (2018) Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, 2018, Published by Elsevier Ltd. https://doi.org/10.1016/j.imu.2017.12.006.

[2] Su Su Maw, Chiyori Haga (2019) Effect of a 2-hour interval between dinner and bedtime on glycated hemoglobin levels in middle-aged and elderly Japanese people: a longitudinal analysis of 3-year health check-up data , BMJ Nutrition, Prevention & Health, 2019; 2:1–10, https://doi:10.1136/bmjnph-2018-000011.

[3] Diabetes Care (2016) Jan; 39(Supplement 1): S13-S22. https://doi.org/10.2337/dc16-S005.

[4] Deepti Sisodiaa, Dilip Singh Sisodiab (2018) Prediction of Diabetes using Classification Algorithms, International Conference on Computational Intelligence and Data Science (ICCIDS 2018). https://doi:10.1016/j.procs.2018.05.122.

[5] Neha Prerna Tiggaa, Shruti Garga (2019) Prediction of Type 2 Diabetes using Machine Learning Classification Methods, International Conference on Computational Intelligence and Data Science (ICCIDS 2019). https://doi:10.1016/j.procs.2020.03.336.

[6] Changsheng Zhua, Christian Uwa Idemudiaa, Wenfang Fengb (2019) Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, Informatics in Medicine Unlocked 17 (2019) 100179. https://doi.org/10.1016/j.imu.2019.100179.

[7] Talha Mahboob Alama,∗, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb,Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad Awais Malikb, Muhammad Mehdi Razab,Salman Ibrarb, Zunish Abbasd, (2016) A model for early prediction of diabetes, Informatics in Medicine Unlocked 16 100204. https://doi.org/10.1016/j.imu.2019.100204.

[8] American Diabetes Association (2016) Classification and diagnosis of diabetes. Sec. 2. In Standards of Medical Care in Diabetes 2016. Diabetes Care 2016; 39 (Suppl. 1): S13–S22.
[9] Aakansha Rathore, Simran Chauhan, Sakshi Gujral (2017) Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women, International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017.
[10] YoichiHayashi n, ShonosukeYukita (2016) Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type2 diabetes mellitus in the Pima Indian dataset, Informatics in Medicine Unlocked 2(2016) 92–104. http://dx.doi.org/10.1016/j.imu.2016.02.001.
[11] Vandana Rawat, Suryakant Suryakant (2019) A Classification System for Diabetic Patients with Machine Learning Techniques. International Journal of Mathematical, Engineering and Management Sciences, 2019, 4 (3), pp.729 – 744. https://10.33889/IJMEMS.2019.4.3-057. hal-02331050.
[12] Anand Kumar Srivastava,Yugal Kumar,Pradeep Kumar Singh (2020) A Rule-Based Monitoring System for Accurate Prediction of Diabetes Monitoring System for Diabetes, International Journal of E-Health and Medical Communications, Volume 11, Issue 3 ,July-September 2020.
[13] Dongmei Pei, Tengfei Yang, Chengpu Zhang (2020) Estimation of Diabetes in a High-Risk Adult Chinese Population Using J48 Decision Tree Model, Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy,2020:13 4621-4630.
[14] A.Tharwat, Applied Computing and Informatics (2018). https://doi.org/10.1016/j.aci.2018.08.003.
[15] Priya. M, M. Karthikeyan (2019) Data Mining Technique for Diabetes Diagnosis using Classification Algorithms, International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-4, November 2019.
[16] Najmeh Hosseinpour, Saeed Setayeshi, Karim Ansari-asl, Mohammad Mosleh (2012) Diabetes Diagnosis by Using Computational Intelligence Algorithms, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, December 2012.
[17] Prema N S, Varshith V, Yogeswar J (2019) Prediction of Diabetes using Ensemble Techniques, International Journal of Recent Technology and Engineering (IJRTE), Volume-7, Issue-6S4, April 2019.
[18] Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515. https://doi:10.3389/fgene.2018.00515.
[19] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore (2014) Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, BioMed Research International,Volume 2014, Article ID 781670, 11 pages. https://dx.doi.org/10.1155/2014/781670.

[20] Gaganjot Kaur, Amit Chhabra (2014) Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications (0975 – 8887), Volume 98 – No.22, July 2014.
[21] Dr. D. Ashok Kumar and R. Govindasamy (2015) Performance and Evaluation of Classification Data Mining Techniques in Diabetes, International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015, 1312-1319.
[22] Mr. R. Sengamuthu, Mrs. R. Abirami, Mr. D. Karthik (2018) Various Data Mining Techniques Analysis to Predict Diabetes Mellitus, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 05 | May-2018.
[23] Komal Patil, Dr. S. D. Sawarkar, Swati Narwane (2019) Designing a Model to Detect Diabetes using Machine Learning, International Journal of Engineering Research & Technology (IJERT), Vol. 8 Issue 11, November-2019.
[24] Leon Kopitar, Primoz Kocbek, Leona C ilar, Aziz Sheikh & Gregor Stiglic (2018) Early detection of type 2 diabetes mellitus using machine learning-based prediction models, Scientific Reports. https://doi.org/10.1038/s41598-020-68771-z.
[25] Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram (2010) Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection,International Journal of Information Technology and Knowledge ManagementJuly-December 2010, Volume 2, No. 2, pp. 271-277.
[26] A. Mary Posonia,S. Vigneshwari,D. Jamuna Rani (2020) Machine Learning based Diabetes Prediction using Decision Tree J48, Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020] IEEE Xplore Part Number: CFP20M19-ART.
[27] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, and Basant Tiwari (2022), A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques, Hindawi Journal of Healthcare Engineering, Volume 2022, Article ID 1684017, 10 pages, https://doi.org/10.1155/2022/1684017
[28] Rashedur M. Rahman, Farhana Afroz (2013) Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis,Journal of Software Engineering and Applications, 2013, 6, 85-97. http://dx.doi.org/10.4236/jsea.2013.63013 Published Online March 2013, https://www.scirp.org/journal/jsea.
[29] Md. Aminul Islam, Nusrat Jahan (2017), Prediction of Onset Diabetes using Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887) Volume 180 – No.5, December 2017.