sensors

MDPI

COMPARITIVE MACHINE LEARNING APPROACH: FOLLOW UP STUDY ON TYPE2 DIABETES PREDICTIONS BY CROSS VALIDATION METHODS

GOPI BATTINENI[1]*, GETUGAMO SAGARO[1], NALINI CHINTALAPUDI[1], FRANCESCO AMENTA [1, 2],
SYED KHASROW TAYEBATI[1]

1. E-Health and Telemedicine Centre, School of Pharmaceutical Sciences and Health Products, University of Camerino, Italy

2. Studies and Research Department, International Medical Radio Center Foundation (C.I.R.M.), Rome, Italy

Corresponding Author*
Dr. Gopi Battineni, gopi.battineni@unicam.it, +39-3331728206.

ABSTRACT

Background

Diabetes is the common chronic diseases that are on the rise, and it had become one of the leading causes for deaths. By early prediction of diabetes, we can improve the healthy lifestyle of patients with diabetes. Currently, many studies were adopting machine learning (ML) techniques for predicting and diagnosing this disease, and in this study we conducted experiments to predict diabetes in Pima Indian females with particular ML classifiers which were exposed to cross validation methods

Methods

Pima Indian diabetes dataset (PIDD) with 768 female patients were considered. Different data mining operations were performed to conduct comparative analysis of four different ML classifiers of Naïve Bayes, J48, Logistic Regression (LR), and Random Forest (RF) was adopted. All the models were exposed to different cross-validations (K=5, 10, 15&20) values, and performance measures such as accuracy, precision, F-score, recall, and AUC were calculated for each model.

Results

Results were mentioned that LR produces the highest accuracy (0.77) for all 'k' values. When k=5, the accuracy of J48, NB, RF were found as 0.71, 0.76, 0.75. Alternatively, for k=10, accuracies of J48, NB, RF were found as 0.73, 0.76, 0.74, and for k=15, 20, the accuracy of NB found as 0.76. Accuracies of J48, RF recorded as 0.76 when k=15 and 0.75 when k=20. In addition to accuracy, other parameters such as precision, f-score, recall, and AUC were considered.

Conclusion

The present study on PIDD was trying to propose an optimized ML model, especially with cross-validation methods. In addition to accuracy, conclusions were made on AUC values, and with AUC LR (0.83), RF (0.82), and NB (0.81) are ranked as three best models to predict whether the patient diabetic or not.

Keywords: Machine Learning (ML), Diabetes, PIDD, Accuracy, Model validation

1. INTRODUCTION

Diabetes is one of the common chronic diseases that exposed when the pancreas does not produce either enough insulin considered as type1 diabetes or when patient body does not effectively utilize the insulin treated as type 2 diabetes. In addition, hyperglycemia or raised blood sugar is a common effect of uncontrolled diabetes. Over the time, it can leads to severe damage to the nerves and blood vessels [1]. Sometimes, it might produce critical health problems like coronary illness, visual impairment, and kidney failures [1], [2]. Therefore, it was becoming one of the driving forces for death rates in the present world. Therefore, to enhance patient life expectancy and improve a quality lifestyle, it is crucial to do early detection of diabetes [3].

Machine learning (ML), is an application of artificial intelligence (AI) that can provide the ability for computers about self-learning and perform statistical analysis without having the human interactions [4]. These algorithms and models were becoming comprehensive and increasingly reliable over all types of industries. Many research groups were trying to adopt them in medical industries, especially for diagnostics, disease predictions [5], drug discovery, and clinical trials [6].

The machine learning process starts with data (either structured or unstructured) from the different sources. The next step is data preparation or data preprocessing is to fix issues related with the interest of data selection. Data preprocessing is the data mining method that involves converting original or raw data into an understandable format[7]. Once the data has ready, model start to test different trained data sets to calculate accuracy or perform statistical algorithms is known as model validation [8]. Ultimately, model optimization or model improvement was done by hyper parameters tuning for final validation to perform prediction, and classification (Refer Figure 1).

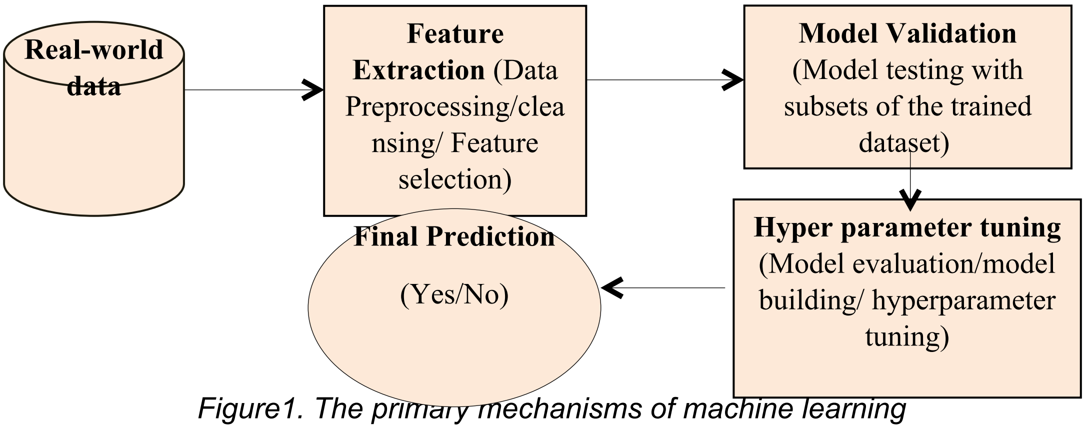

Real-world data
Feature Extraction (Data Preprocessing/cleansing/ Feature selection)
Model Validation (Model testing with subsets of the trained dataset)
Hyper parameter tuning (Model evaluation/model building/ hyperparameter tuning)
Final Prediction (Yes/No)

Figure1. The primary mechanisms of machine learning

In healthcare industries, large amounts of patient data and medical knowledge stored in databases demanding to development of new tools and technologies for data analysis and classification. Currently, ML algorithms are operated to automatic analysis of high dimensional medical data. Dementia forecasting [9], cancer tumor identification [10], diabetes predictions [11], and radiotherapy [12] were some of the examples of ML in medicine.

According to WHO reports, 425 million people globally having diabetes [13]. Many researchers have conducted studies on diagnosing and early prediction of diabetes. Outcomes were suggested that risk factors associated with type2 diabetes were family history, hypertension, unhealthy diet, lack of physical activities, and overweight. Unfortunately, this rate is getting higher in females, especially at pregnancy time and it is because of low percentage of insulin consumption, high cholesterol level, or rise in blood pressure [13], [14]. With the presence of contemporary technologies in computer science, studies have shown that by the employment of computer skills and data mining algorithms, cost-effective and efficient techniques could be developed for a diabetes diagnosis.

Many studies were conducted prediction analysis using data mining algorithms for diagnosing diabetes. For instance, in the study of [15], researchers utilize support vector machines (SVM) for the diagnosis of diabetes mellitus and achieve a prediction accuracy about 94%. In the research [16], scholars used J48 decision trees, RF, and neural networks were used, and results mentioned as RF was producing the highest accuracy

80.4% in diabetic patient classifications. Another study [17], proposed a model for forecast the likelihood of diabetes mentioned that Naïve Bayes (NB) produced the 76.3% accuracy that was higher than J48, and SVM. In ref [18], accuracy analysis was done over different ways of data preprocessing, and parameter modification to improve the accuracy. Outcomes mentioned deep neural networks (DNN) with cross-validation (K=10) generated 77.86% accuracy in diabetes identification.

In this study, we developed a classification model for type2 diabetes in Pima Indian females. Adoption of four classification ML algorithms like J48 decision trees, NB, RF, and Logistic regression (LR) in diabetes detection of female patients. We approach the cross-validation (CV) techniques to train the different ML models for different test data sets. In addition, ranking of each algorithm was decided depending on performance parameters such as accuracy, precision, recall and F-scores.

2. METHODS AND MATERIALS

A dataset with Pima Indian female patients of at least twenty-one year's age that has been taken from the UCI machine learning repository. This dataset is originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases. It is open-source one with objective of diagnostic diabetic prediction of 768 instances classified into two classes of tested positive (class variable: 1), and tested negative (class variable: 0) with eight different risk factors: number of times pregnant (preg); plasma glucose concentration a 2 hours in an oral glucose tolerance test (plus); diastolic blood pressure (pres); triceps skinfold thickness (skin); 2-Hour serum insulin (insu); body mass index (mass); pedigree function (Pedi); and age (age) as in Table 1.

Table1: Statistical report of Pima Indian diabetes dataset (PIDD)

Attribute Number	*Risk factor*	*Variable Type*	*Range (min-max)*
1	Number of times pregnant	Integer	0-17
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Integer	44-199
3	Diastolic blood pressure (mm Hg)	Integer	24-122
4	Triceps skinfold thickness (mm)	Integer	7-99
5	2-Hour serum insulin (mu U/ml)	Integer	14-846
6	Body mass index (weight in kg/(height in m)^2)	Real	18.2-67.1
7	Diabetes pedigree function	Real	0.07-2.42
8	Age (years)	Integer	21-81
9	Class	Binary	1-Tested Positive (268) 0-Tested Negative (500)

Investigation of dataset was done by WEKA 3.8, which is an open-source tool to perform data mining and machine learning operations. Later PIDD was exposed to data preprocessing steps were conducted to avoid unbalanced datasets (Refer Figure2). To identify primary diabetic causes, we generated pruned decision tree with model training that has done by the dataset after the removal of missing values.

Data sampling techniques were performed to manage imbalanced datasets. In further, the model classifiers were exposed to different cross-validations (k- folds). Cross-validation (CV) is defined as a model training method that can used to assess the prediction accuracy for judging how model performance was affected by test data [19].

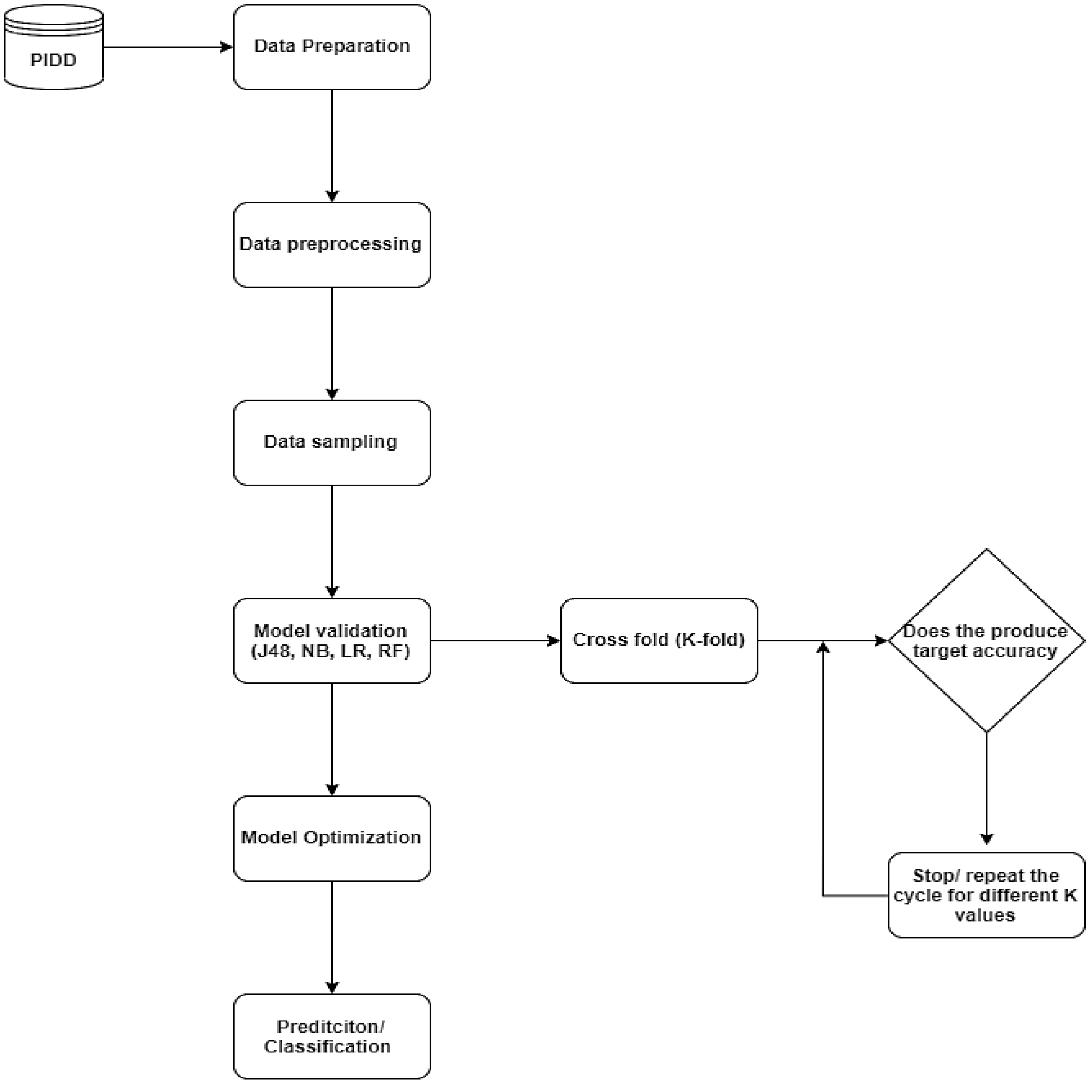

PIDD
Data Preparation
Data preprocessing
Data sampling
Model validation (J48, NB, LR, RF)
Cross fold (K-fold)
Does the produce target accuracy
Stop/ repeat the cycle for different K values
Model Optimization
Preditciton/ Classification

Figure2. Approached methodology

2.1. Data Sampling:

Data sampling is a one of the machine learning technique to fix imbalanced datasets into balanced ones. There are two sampling techniques, either oversampling on the minority class instances or under-sampling on the majority class instances. Different forms of PIDD datasets with statistical values for each attribute were found in Table 2.

Table2. Statistics of original and different trained sets (where SD: Standard deviation)

Statistics	Dataset	Preg	Plas	Pres	Skin	Insu	BMI	Pedi	Age
Count	Original	768	768	768	768	768	768	768	768
	Preprocess	392	392	392	392	392	392	392	392
	Under-sampling	536	536	536	536	536	536	536	536
	Oversampling	1036	1036	1036	1036	1036	1036	1036	1036

Mean	Original Preprocess Under-sampling Oversampling	3.84 3.30 1 4 4.08 4	121.6 122.62 126.228 126.123	72.40 70.66 69.095 69.593	29.15 29.14 20.403 20.818	155.54 156.05 84.981 84.894	32.45 33.08 32.553 32.765	0.47 2 0.52 3 0.48 8 0.49 4	33.241 30.86 33.944 34.2
SD	Original Preprocess Under-sampling Oversampling	3.37 3.21 1 3.46 4 3.34 9	30.43 30.86 33.335 32.443	12.09 12.49 20.378 19.378	8.79 10.51 16.515 16.062	85.02 118.84 124.84 121.33	6.87 7.028 7.877 7.522	0.33 1 0.34 5 0.35 1 0.33 2	11.76 10.201 11.684 11.43
Min	Original Preprocess Under sampling Over sampling	0 0 0 0	0 56 0 0	0 24 0 0	0 7 0 0	0 14 0 0	0 18.2 0 0	0.07 0.08 5 0.07 8 0.07 8	21 21 21 21
Max	Original Preprocess Under-sampling Oversampling	17 17 17 17	199 198 199 199	122 110 114 122	99 63 99 99	846 846 846 846	67.1 67.1 67.1 67.1	2.42 2.42 2.42 2.42	81 81 81 81

2.2. Cross-Validation methods

The great challenge in machine learning techniques was model validation on trained data, and sometimes, this could be hard to assess whether the model will work correctly for the trained data. Therefore, to ensure adopted model is producing specific patterns without having much noise [20], data scientists were using cross-validation (CV) techniques. It is a popular method in ML environment because of easy understanding of low biased model estimation when compare with other methods.

Model validation was done with the K-fold cross-validation techniques. Primarily, original dataset split into k folds to perform training on the total data subsets and remaining k-1 subsets are combined to form a trained dataset. Thus, original data was randomly separated into 'k' folds of like k_1, k_2, …….., $k_{n.}$ The testing and training process performed in 'k' times. In iteration i, subset K_i is work as the test data, and remaining subsets combined to form as trained data. For example, in the first iteration, if subset k_1 is served as test data, then remaining subsets k_2,….., k_n were combined for model training, and this process is repeated for the rest of the k values. The optimal value of 'k' was 5 or 10, but not restricted to other values. In this study, we perform model iterations on k=5, 10, 15, and 20. If 'k' is higher, the difference between sizes of training dataset and resampling datasets gets smaller. If the difference in data sizes decreases, the model technique bias will become low.

2.3. Naïve Bayes (NB)

Naïve Bayes (NB) is a probabilistic ML model approach can used for classification. It depends on the assumption that it is rarely valid in practical learning problems. Besides, it allocates a probability to target group in the classification analysis based on feature extraction [21]. Generally, this algorithm is easy and fast to predict the test data and produce a good performance in multi-class predictions. In the case of categorical input variables, NB performs well when compared to numerical values. As of this, we adopted NB for performing diabetic classification.

The representation of Bayes theorem was done by equation 1

P (c/ X) = -------------- (1)____

The probability of ‘c’ is happening given that ‘X’ occurrence.
Here, P (c/X) = target class’s posterior probability,
P (X/c) = predictor class’s probability,
P(c) = class ‘c’s probability is true,
P(X) = predictor’s prior probability.

2.4. Logistic Regression (LR)

LR is a classification algorithm used to allocate observations to the discrete set of classes. Basically, it is classified into the binary, multi, and ordinal level type. Instead of the relationship between the non-continuous attributes, LR permits the prediction of discrete variables by the mixed values of continuous and discrete predictors [22]. Because of easy implementation and efficient to train, we selection of LR technique was done in this study.

Logistic regression is mathematically written as multiple linear regression function (refer equation I) by

------ (I)

The following example will explain a simple logistic binary function. As discussed, two target diabetic groups (tested positive- ‘1’ or tested negative-‘0’),

Hypothesis W = AX+B --------- (II), and

H (x) = sig (W) ------- (III)

If ‘W’ goes to infinity, output prediction will be tested positive, and if ‘W’ goes to negative infinity, output prediction will be tested negative (Figure 3).

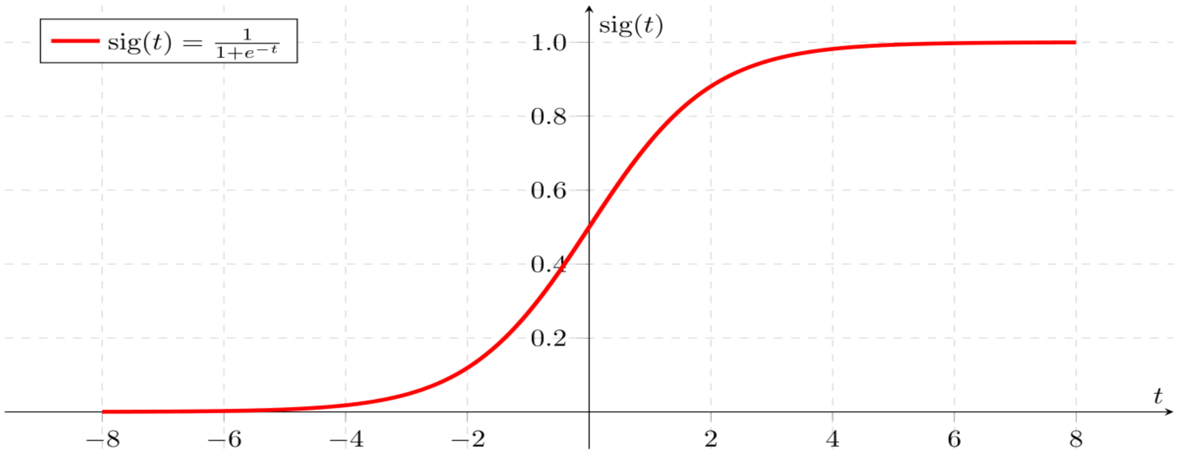
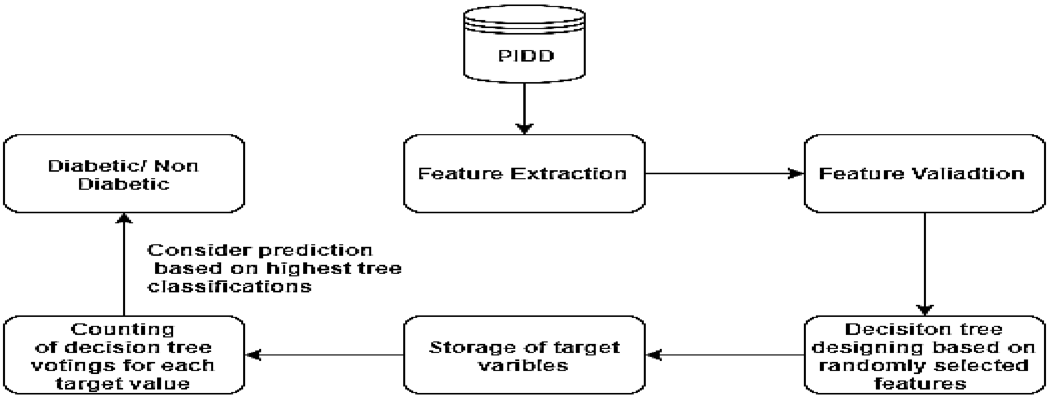


Figure3. Simple binary logistic regression representation (where sig (t) sigmoid activation function)

2.5. Random Forest (RF)

Because of utilizing feature selection methods, RF algorithms were quick in learning and produce highest classification accuracy on large databases. The reason behind this was usage of tree-based systems by RF, usually positions by how well they improve the virtue of the node. This means a decrease in impurity over all trees, which called Gini impurity [23]. In RF, the first step is to feature extraction need to be done from the test data group. Later, we need to validate the test features by the randomly created decision trees (Refer Figure4). For example of PIDD, if model produces 50 random trees and each tree will predict two different outcomes for the same test group. If 30 trees were predicted as ‘tested positive' and 20 trees were predicted as ‘tested negative,' then RF algorithm returns ‘tested positive’ as the predicted target.


PIDD
Feature Extraction
Feature Valiadtion
Decisiton tree designing based on randomly selected features
Storage of target varibles
Counting of decision tree votings for each target value
Consider prediction based on highest tree classifications
Diabetic/ Non Diabetic

Figur4. RF procedure flow chart representation

2.6. J48

J48 or decision tree algorithms used to calculate the feature behavior for different test groups which can used to generate the rule for predictions of the dependent variables. Consequently, by using the J48 algorithm, we can understand the critical distribution of instances. Besides, it helps to identify missing attributes, and used as a precision tool when data over fitting occurred [24]. However, major challenge in the decision trees was identifying the root node attribute, the attribute selection measures in decision trees were done in two ways (1) Information gain, and (2) Gini Index.

Suppose X is given a set of instances, A is an attribute, X_x is the subset of X with A=X. Also, value (A) is set total possible values of A, and then information gain defined in equation 1,

--------- (1)

Gini index (GI) is a parameter that helps to calculate how often randomly selected instances could be incorrectly classified. The formula for calculating GI is defined in equation 2.

---------------- (2)

2.7. Performance Measures

Model performance was decided by accuracy, precision, recall, and F-measure values. In general, accuracy is defined as the percentage of correctly classified instances from total instances. All performance measures, along with formulation and definitions, can be found in table 3.

Table3: Definition and Formulation of Accuracy Measures (Where TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative)

Parameter	**Definition**	**Formulation**
Accuracy	rate of correctly classified instances from total instances	
PRECISION (P)	Rate of correct predictions	
RECALL (R)	True Positive Rate	
F-Measure	Used to measure the accuracy of the experiment	

3. RESULTS

Exclusion of oversampling and under sampling datasets was done due to issues of model over fitting or left important instances from the trained datasets. We exposed each model for validation of trained dataset on different k values

3.1. Pruned decision tree

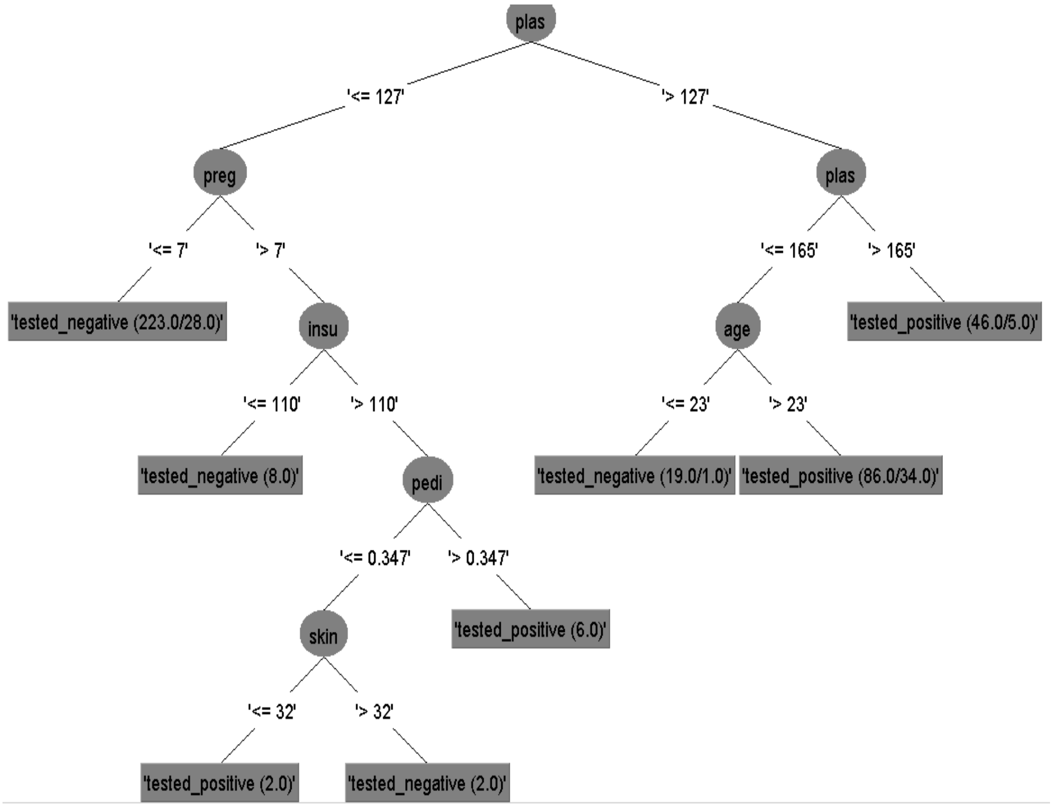


Figure5. Pruned decision tree

To avoid data over fitting, we conducted a pruning technique associated with decision trees. We exposed the model with the dataset after missing instances removal and conducted a J48 decision tree pruning with k=5. The pruning tree outcome with glucose value as a central node can see in Figure 5. It is evident that glucose tolerance has the highest information gain, so it is the highest risk factor for getting diabetes. Other risk factors were getting several times pregnancy, the release of high insulin, and pedigree function. Pregnant women usually not exposed to physical exercises that might make them high weight gain, and sometimes, this can increase the problem of getting type 2 diabetes.

3.2. Confusion matrix

Generally, a confusion matrix is defined as the table, which mostly used to describe the performance of model classifiers[25]. We performed a four-machine learning model simulation in this study to analyze the accurate prediction of the test class (See Table 4).

Table 4: confusion matrix of different classifier models

A	B	<-- classified as	Model
427	73	A = Tested negative	Naïve Bayes
122	146	B = Tested positive	
450	50	A = Tested negative	Logistic Regression
129	139	B = Tested positive	
431	69	A = Tested negative	Random Forest
118	150	B = Tested positive	
427	73	A = Tested negative	J48
122	146	B = tested positive	

3.3. Model classification

We conduct experiments on four different classification models (J48, NB, RF, and LR) to diagnose the patient, whether diabetic or non-diabetic. Table5 gives hyper parameters of different model classifier that trained to classify diabetes of Pima Indian woman patients. Performance measures validate all models that exposed to different cross-validations to perform optimization. The performance of four models evaluated based on parameters such as accuracy, recall, precision, AUC (area under the curve), and F-scores was found in Table6.

Table5. Hyper parameters of different classifiers (here C: Pruning confidence and 'R' – R squared value)

Number	Model	Tuning Parameters
1.	J48	C = 0.25
2.	NB	-
3.	RF	Number of trees- 100, Number of features to construct each tree -4, and out of bag error: 0.237
4.	LR	R =1.0E-8

Table6. Performance measures of different models classifiers (where k = 5, 10, 15&20)

K-values	Classifier	Accuracy	Precision	Recall	F-Score	AUC
K=5	J48	0.71	0.71	0.71	0.71	0.72
	NB	0.76	0.76	0.76	0.76	0.81
	RF	0.75	0.75	0.75	0.75	0.82
	LR	0.77	0.77	0.77	0.76	0.83
K=10	J48	0.73	0.73	0.73	0.73	0.75
	NB	0.76	0.75	0.76	0.76	0.81
	RF	0.74	0.74	0.74	0.74	0.81
	LR	0.77	0.76	0.77	0.76	0.83
K=15	J48	0.76	0.75	0.76	0.76	0.74

	NB	0.76	0.75	0.76	0.75	0.81
	RF	0.76	0.76	0.76	0.76	0.82
	LR	0.77	0.77	0.77	0.76	0.83
K=20	J48	0.75	0.74	0.75	0.74	0.74
	NB	0.76	0.75	0.76	0.75	0.81
	RF	0.75	0.74	0.75	0.74	0.82
	LR	0.77	0.77	0.77	0.76	0.83

4. DISCUSSION

Early prediction of diabetes will help to provide a healthy lifestyle for diabetic patients. Present study is aiming to propose an optimized machine learning algorithm to classify and diagnose diabetic patients with k-fold validation methods. Randomly selected four supervised ML classifiers were used for the development of different models in diagnosis of diabetes especially on Pima Indian female population. Cross-validation methods were adopted for different k values to perform model improvement. But, by producing similar accuracies, final rankings of each model assessed by receiver optimistic curve (ROC) rates.

ROC is the visualizing tool of the binary classifiers performance. It is generated by plotting false positive rate (x-axis) against the true positive rate (y-axis) to decide correct threshold value [26]. The area under the curve (AUC) is the rate of accurate model classification, and the typical range of AUC was 0.5-1.0. If AUC is near to value 1, the model was done accurate instance classification, and proper optimization was done [27]. Four different machine learning algorithms have been developed for various k values to predict whether a patient is diabetic or non-diabetic. Dataset was splitting into 'k' subsets to perform training and testing with k number of times. All preliminary analysis was performed with Weka studio.

From Table 6, it is found that the LR model is producing a high accuracy of 0.77 when compared with others. To avoid overfitting and under fitting issues, tenfold cross-validation is considered. The highest accuracy is achieved when trained data exposed to k=10, NB generated 0.76, and the remaining J48, RF is producing accuracies with 0.73, 0.74. Recall or sensitivity defines the rate of correctly predicted diabetic patients. For LR, it is found as 0.77, and for RF, J48, and NB, it is 0.74, 0.73, and 0.76. The precision of NB and J48 found as 0.75 and 0.73, and for RF, LR it is 0.74, 0.76. F scores of J48, NB, RF, and LR were reported as 0.73, 0.76, 0.74, and 0.76. Besides, we also calculate AUC to measure the performance of the four models. The AUC of J48, NB, RF, and LR was reported as 0.75, 0.81, 0.81, and 0.83.

From the above studies, it understands that all four classifiers were producing similar prediction accuracies with little margins. However, LR reported the highest accuracy, and J48 reported low accuracy when compared with others. Ultimately, LR, NB, and RF are considered as the three best models to predict patient as diabetic or not. Further, for K (5, 10, and 20), accuracy, precision, recall, and f-scores of NB were higher compared to RF. However, for K=15, precision and F-scores of RF was reported higher than NB. Accuracy is not only the parameter that decides model optimization, and the main limitation of using accuracy as a key performance metric; it does not work well in datasets having severe class imbalances. If we carefully observed our diabetic dataset (table 1), 500 instances found tested negative, and 268 instances found tested positive that producing imbalance ratio 1.87.

Therefore, along with the mentioned four parameters, it is also important to consider AUC values. It can be seen that the AUC values of NB (Figure 6.1), LR (Figure 6.2) were 0.81, 0.83 and for RF (Figure 6.3) it is found as 0.82 (k=5, 15, &20), and 0.81 (k=10). However, J48 produces low AUC value (0.72) while comparing it with others that represent in Figure 6.4. By summing all these results, when we tried to assign rankings for four classifiers based on performance values to produce model optimization followed as LR>RF>NB>J48.

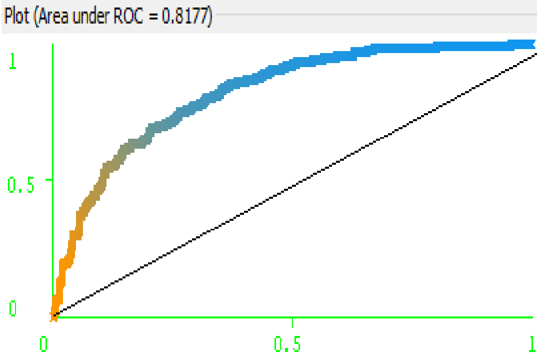


Figure 6.1 AUC curve of NB

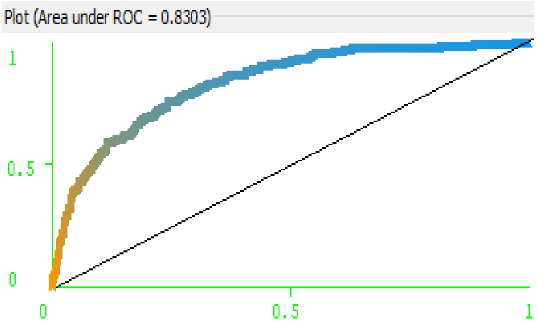


Figure 6.2 AUC curve of LR

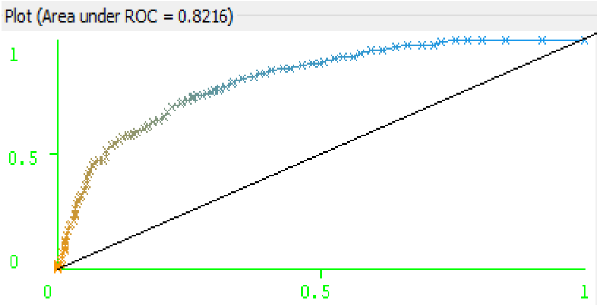


Figure 6.3 AUC curve of RF

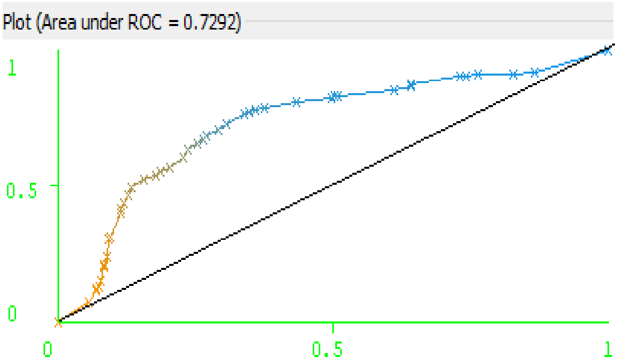


Figure 6.4 AUC curve of J48

5. CONCLUSIONS

Diabetes is one of the critical chronic diseases that affect most of the world population. Doctors are proposing preventive methods to get a cure from this. Also, with the latest trends in machine intelligence, it is overcoming by knowing the reasons before of having diabetes. We developed four binary classifier models, such as NB, J48, LR, and RF. Each model exposed to different cross-validation methods subject to different 'k' values. Performance assessment was done based on various parameters such as accuracy, precision, recall, F-scores, and AUC. Preliminary outcomes suggested that all models achieved good results, but the LR model producing the highest accuracy 0.77 for all k-values and J48 produces relatively low accuracy when compared with others. The ranking assignment was done on every algorithm by considering other parameters along with accuracy. Therefore, it is evident that LR, NB, RF are the three best models to predict patient as to whether diabetic or not.

The main limitation of this study, we only consider conventional ML classifiers even it addressed better performance in predicting diabetes. There should be mandatory to conduct more analysis of diabetic forecasting with the help of other contemporary methods in future research.

Conflict of Interest: None

Acknowledgments

Gopi Battineni, Getugamo Sagaro, Nalini Chintalapudi Ph.D. bursaries were supported in part by the University of Camerino.

Author Note

We are certifying that the manuscript is not under review by any journal. All authors are strictly read and validated the final copy. GB*: Design and perform the experiments. Analyze the methods and wrote the manuscript. CN& GS: Validate the results and contribution on literature review, SKT & FA: Conclusion and final manuscript revision

ORCID

Gopi Battineni: https://orcid.org/0000-0003-0603-2356

REFERENCES

[1] S. R. K. Seshasai et al., "Diabetes mellitus, fasting glucose, and risk of cause-specific death," N. Engl. J. Med., 2011.

[2] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," The Lancet. 2017.

[3] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Informatics, 2019.

[4] Y. Baştanlar and M. Özuysal, "Introduction to machine learning," Methods Mol. Biol., 2014.

[5] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)," Informatics Med. Unlocked, 2019.

[6] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine. 2019.

[7] D. P. Methods, "Data Preprocessing Techniques for Data Mining," Science (80-.)., 2011.

[8] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," J. Mach. Learn. Res., 2010.

[9] S. Mathotaarachchi et al., "Identifying incipient dementia individuals using machine learning and amyloid imaging," Neurobiol. Aging, 2017.

[10] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, and H. J. W. L. Aerts, “Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer,” Front. Oncol., 2015.

[11] N. Nirala, R. Periyasamy, B. K. Singh, and A. Kumar, “Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine,” Biocybern. Biomed. Eng., 2019.

[12] M. L. Giger, “Machine Learning in Medical Imaging,” J. Am. Coll. Radiol., 2018.

[13] N. G. Forouhi and N. J. Wareham, “Epidemiology of diabetes,” Medicine (United Kingdom). 2019.

[14] N. G. Forouhi, A. Misra, V. Mohan, R. Taylor, and W. Yancy, “Dietary and nutritional approaches for prevention and management of type 2 diabetes,” BMJ, 2018.

[15] N. Barakat, A. P. Bradley, and M. N. H. Barakat, “Intelligible support vector machines for diagnosis of diabetes mellitus,” IEEE Trans. Inf. Technol. Biomed., 2010.

[16] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Front. Genet., vol. 9, no. November, pp. 1–10, 2018.

[17] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” in Procedia Computer Science, 2018.

[18] S. Wei, X. Zhao, and C. Miao, “A comprehensive exploration to the machine learning techniques for diabetes identification,” in IEEE World Forum on Internet of Things, WF-IoT 2018 - Proceedings, 2018.

[19] S. Watanabe, “Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory,” J. Mach. Learn. Res., 2010.

[20] C. Bergmeir and J. M. Benítez, “On the use of cross-validation for time series predictor evaluation,” Inf. Sci. (Ny)., 2012.

[21] T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” Int. J. Comput. Sci. Appl. ISSN 0974-1011, 2013.

[22] A. I. Schein and L. H. Ungar, “Active learning for logistic regression: An evaluation,” Mach. Learn., 2007.

[23] B. H. Menze et al., “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” BMC Bioinformatics, 2009.

[24] S. Tsang, B. Kao, K. Y. Yip, W. S. Ho, and S. D. Lee, “Decision trees for uncertain data,” IEEE Trans. Knowl. Data Eng., 2011.

[25] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manag., 2009.

[26] D. J. Lingenfelter, J. A. Fessler, C. D. Scott, and Z. He, “Predicting ROC curves for source detection under model mismatch,” in IEEE Nuclear Science Symposium Conference Record, 2010.

[27] J. Huang and C. X. Ling, “Using AUC and accuracy in evaluating learning algorithms,” IEEE Trans. Knowl. Data Eng., 2005.