machines

MDPI

Article

Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods

Gopi Battineni [1,*], Getugamo Sagaro [1], Chintalapudi Nalini [1], Francesco Amenta [1,2] and Seyed Khosrow Tayebati [1]

[1] E-Health and Telemedicine Centre, School of Pharmaceutical Sciences and Health Products, University of Camerino, 62032 Camerino, Italy

[2] Studies and Research Department, International Medical Radio Center Foundation (C.I.R.M.), 00144 Rome, Italy

* Correspondence: gopi.battineni@unicam.it; Tel: +39-3331728206.

Received: 10 September 2019; Accepted: 23 October 2019; Published: 5 December 2019

Abstract: (1) Background: Diabetes is a common chronic disease and a leading cause of death. Early diagnosis gives patients with diabetes the opportunity to improve their dietary habits and lifestyle and manage the disease successfully. Several studies have explored the use of machine learning (ML) techniques to predict and diagnose this disease. In this study, we conducted experiments to predict diabetes in Pima Indian females with particular ML classifiers. (2) Method: A Pima Indian diabetes dataset (PIDD) with 768 female patients was considered for this study. Different data mining operations were performed to a conduct comparative analysis of four different ML classifiers: Naïve Bayes (NB), J48, Logistic Regression (LR), and Random Forest (RF). These models were analyzed by different cross-validation (K = 5, 10, 15, and 20) values, and the performance measurements of accuracy, precision, F-score, recall, and AUC were calculated for each model. (3) Results: LR was found to have the highest accuracy (0.77) for all 'k' values. When k = 5, the accuracy of J48, NB, and RF was found to be 0.71, 0.76, and 0.75. For k = 10, the accuracy of J48, NB, and RF was found to be 0.73, 0.76, 0.74, while for k = 15, 20, the accuracy of NB was found to be 0.76. The accuracy of J48 and RF was found to be 0.76 when k = 15, and 0.75 when k = 20. Other parameters, such as precision, f-score, recall, and AUC, were also considered in evaluations to rank the algorithms. (4) Conclusion: The present study on PIDD sought to identify an optimized ML model, using with cross-validation methods. The AUC of LR was 0.83, RF 0.82, and NB 0.81). These three were ranked as the best models for predicting whether a patient is diabetic or not.

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

1. Introduction

Diabetes is a common chronic disease occurring when the pancreas does not produce enough insulin (Type 1 diabetes) or when the patient's body does not effectively utilize the insulin (Type 2 diabetes). Hyperglycemia or raised blood sugar is the common consequence of uncontrolled diabetes. Over time, diabetes can cause severe damage to nerves and blood vessels [1]. Advanced diabetes is complicated by coronary illness, visual impairment, and kidney failure [1,2]. Early detection of the disease can give patients the opportunity to make the necessary lifestyle changes and therefore can improve their life expectancy [3].

Machine learning (ML), is an application of artificial intelligence (AI) that enables computers to self-learn and perform statistical analysis without human interaction [4]. ML algorithms and models are extensively used and have been found reliable for a variety of applications. Researchers have been adopting ML in medicine, especially for diagnosis, disease prediction [5], drug discovery, and clinical trials [6].

The machine learning process starts with structured or unstructured data from different sources. The next step is data preparation or data preprocessing, which involves data selection through a data mining method in which original or raw data is converted into an understandable format [7]. Once the data is ready, the model tests different trained data-sets to calculate accuracy or perform statistical algorithms; this is known as model validation [8]. Model optimization or model improvement is done by hyperparameter tuning for final validation to perform prediction, and classification (Figure 1).

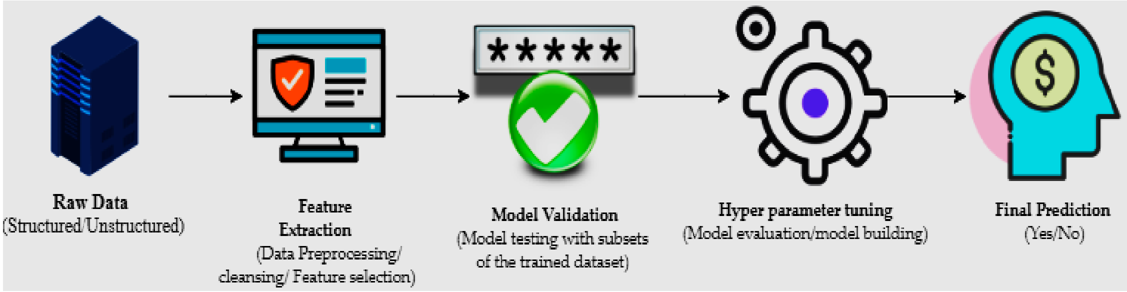
Raw Data
(Structured/Unstructured)
Feature Extraction
(Data Preprocessing/ cleansing/ Feature selection)
*****
Model Validation
(Model testing with subsets of the trained dataset)
Hyper parameter tuning
(Model evaluation/model building)
Final Prediction
(Yes/No)

Figure 1. The primary mechanisms of machine learning.

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

According to World Health Organization (WHO) reports, there are 425 million people in the world with diabetes [13]. Extensive studies on the diagnosis and early prediction of diabetes have shown that the risk factors associated with Type 2 diabetes include family history, hypertension, unhealthy diet, lack of physical activities, and being overweight. Females have a higher tendency to become diabetic (especially during pregnancy), due to low insulin absorption, high cholesterol levels, or a rise in blood pressure [13,14]. Studies have shown that cost-effective and efficient techniques for diagnosing diabetes could be developed by employing computer skills and data mining algorithms.

Several studies conduct prediction analysis using data-mining algorithms to diagnose diabetes. For example, in [15], researchers utilized support vector machines (SVM) for the diagnosis of diabetes mellitus and achieved a prediction accuracy of about 94%. Another work has used J48 decision trees, RF, and neural networks, and has found that RF provides the highest accuracy (80.4%) in diabetic patient classification [16]. Another paper proposed a model to forecast the likelihood of diabetes. This study concluded that Naïve Bayes (NB) had 76.3% accuracy, higher than J48 and SVM [17]. An accuracy analysis conducted on different ways of data preprocessing, and parameter modification was done to improve model precision [18]. The above results revealed that 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 Type 2 diabetes in Pima Indian females. Four classification ML algorithms to detect diabetes in female patients were used: J48 decision trees, NB, RF, and Logistic Regression (LR). Cross-validation (CV) techniques were employed to train the different ML models for varying test data-sets. The ranking of each algorithm was decided based on the performance parameters of accuracy, precision, recall, and F-scores.

2. Methods and Materials

A Pima Indian diabetes dataset (PIDD) with 768 female patients was considered. This dataset, owned by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), contained a tested positive (class variable: 1) and a tested negative (class variable: 0) with eight various risk factors (Table 1).

Table . Statistical report of the Pima Indian diabetes dataset (PIDD).

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

Data investigation was undertaken using WEKA 3.8 [19], which is an open-source tool that can help to perform various data-mining operations. At first, PIDD was exposed to data preprocessing steps to control the unbalanced data-sets (Figure 2).

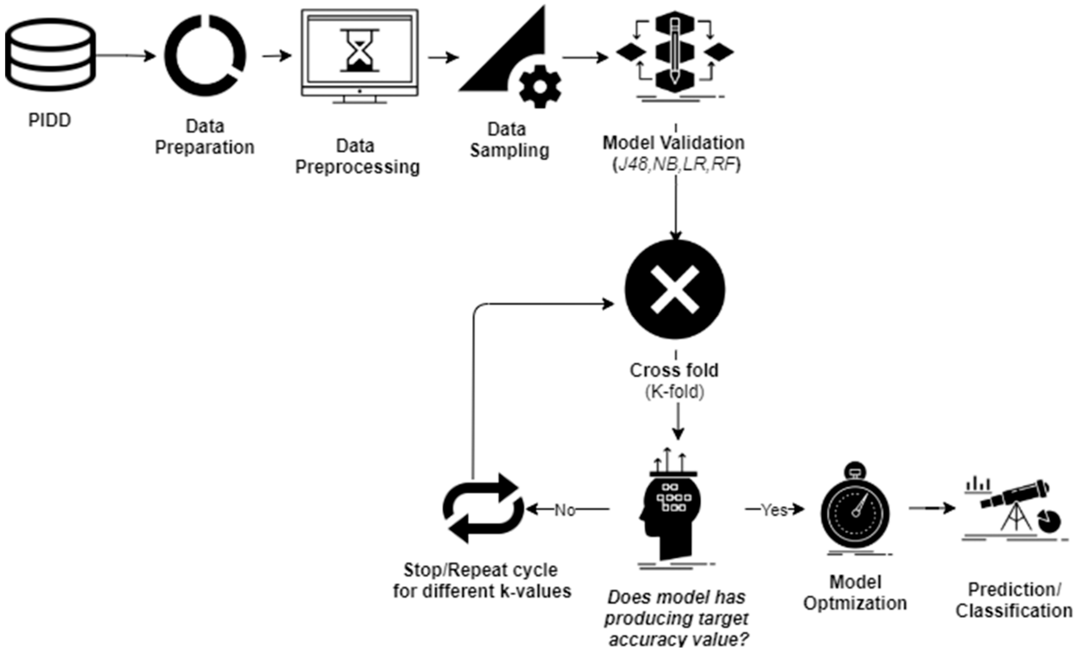
PIDD
Data Preparation
Data Preprocessing
Data Sampling
Model Validation (J48,NB,LR,RF)
Cross fold (K-fold)
Stop/Repeat cycle for different k-values
No
Does model has producing target accuracy value?
Yes
Model Optmization
Prediction/ Classification

Figure 2. Applied methodology.

2.1. Data Sampling

Two data sampling techniques were used to convert imbalanced datasets into balanced datasets: oversampling (on the minority class instances), and under-sampling (on the majority class instances). Different forms of the PIDD dataset with statistical values of each attribute are presented in Table 2.

Table . Statistics of original and different trained sets (where SD: standard deviation, BMI: Body Mass Index).

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	3.84	121.6	72.40	29.15	155.54	32.45	0.472	33.241
	Preprocess	3.301	122.62	70.66	29.14	156.05	33.08	0.523	30.86
	Under-sampling	4	126.228	69.095	20.403	84.981	32.553	0.488	33.944
	Oversampling	4.084	126.123	69.593	20.818	84.894	32.765	0.494	34.2
SD	Original	3.37	30.43	12.09	8.79	85.02	6.87	0.331	11.76
	Preprocess	3.211	30.86	12.49	10.51	118.84	7.028	0.345	10.201
	Under-sampling	3.464	33.335	20.378	16.515	124.84	7.877	0.351	11.684
	Oversampling	3.349	32.443	19.378	16.062	121.33	7.522	0.332	11.43
Min	Original	0	0	0	0	0	0	0.07	21
	Preprocess	0	56	24	7	14	18.2	0.085	21
	Under-sampling	0	0	0	0	0	0	0.078	21
	Oversampling	0	0	0	0	0	0	0.078	21
Max	Original	17	199	122	99	846	67.1	2.42	81
	Preprocess	17	198	110	63	846	67.1	2.42	81
	Under-sampling	17	199	114	99	846	67.1	2.42	81
	Oversampling	17	199	122	99	846	67.1	2.42	81

2.2. Cross-Validation

Cross-validation (CV) is a model training method that can assess prediction accuracy [20]. The biggest challenge in ML is validating the model with trained data. To ensure the adopted model is producing the noise-free model patterns [21], data scientists use CV techniques. Compared with other methods, the CV technique offers the most ease in estimating low bias models, and therefore is one of the most popular techniques in ML algorithms.

In this study, four ML classifiers were employed to conduct different cross-validations. The k-fold CV technique was used to perform model validation. The PIDD was split into 'k' folds to conduct training with test data, and the remaining 'k-1' folds were combined to form trained data. Original data were randomly separated into 'k' folds (k_1,k_2…,k_i), and the model testing was performed by 'k' times. For example, in the first iteration, if subset (k_1) served as test data, then the remaining subsets (k_2,…..,ki) were combined to conduct model training, and this process was repeated for the rest of the 'k' values. Many studies reported that in order to avoid issues associated with imbalanced data-sets, the optimal value for 'k' should be 5 or 10. With the highest (k) values, the difference in trained and sampled data-sets tended to acquire low values. In the present study, model validation was conducted with k = 5, 10, 15, and 20.

2.3. Naïve Bayes (NB)

Naïve Bayes (NB) is a probability-based ML method that can be used as a classification technique. Based on feature extraction, NB produces the probability for target groups in classification [22]. This algorithm quickly and easily predicts the test data and produces better performance values in multi-class predictions. Compared with numerical inputs, NB predicts correctly categorical input values. The Bayes theorem is represented in equation (1) below

$$P\ (c/X) = \frac{(P\ (X|c)\ P(c)}{P(X)} \tag{1}$$

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 into discrete set of classes. It is classified into the binary, multi, and normal level types. LR does not indicate a relationship between non-continuous attributes, but allows the prediction of the discrete variables [23]. It is very easy to implement and quite efficient for training the model.

Logistic regression is mathematically written as a multiple linear regression function (equation 2) by

$$\text{Logit (P)} = \left(\frac{m(x = 1)}{1-(p = 1)}\right) = \beta + \beta 1.x1 + \beta 2.x2 - -\beta i.xm \text{ for } i = 1 \ldots .N \tag{2}$$

The following example represents a simple logistic binary function. As discussed, two target diabetic groups (tested positive- '1' or tested negative-'0') were tested

$$\text{Hypothesis } W = AX{+}B \tag{3}$$

$$H\ (x) = \text{sig}\ (W) \tag{4}$$

If 'W' reaches positive infinity, then the prediction is positive, and if 'W' reaches to negative infinity, then the prediction is negative (Figure 3).

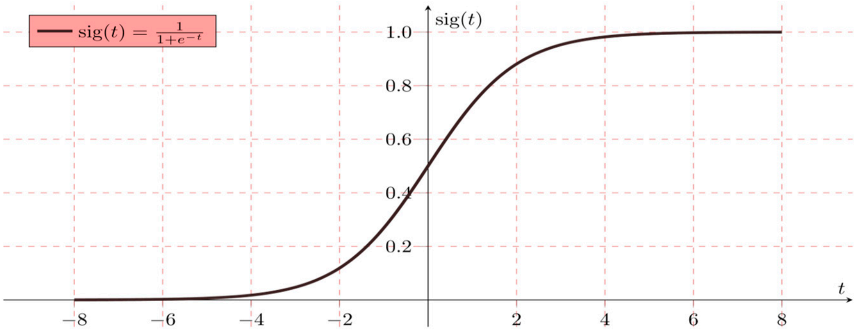

Figure 3. Simple binary logistic regression representation (where sig (t) sigmoid activation function).

2.5. Random Forest (RF)

When feature selection methods are used, the RF algorithm is quick to learn to produce the highest classification accuracy on large databases, because of the tree-based systems used. Generally, these trees are nicely positioned for improving the virtue of the tree node known as the Gini impurity [24]. In RF, feature extraction is conducted from the test data. Thereafter, test features are validated by the randomly generated decision trees (Figure 4). In the example of PIDD, if the model was generated 50 random trees, every tree could predict two different outcomes for the same test group. If 30 trees were predicted (tested positive) and 20 trees were predicted (tested negative), then the RF algorithm returns 'tested positive' as the predicted target.

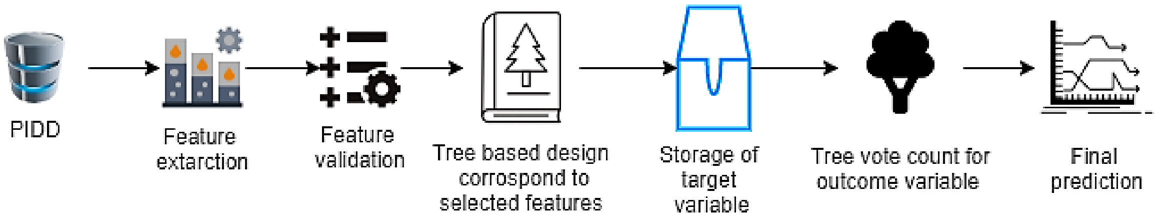
PIDD
Feature extarction
Feature validation
Tree based design corrospond to selected features
Storage of target variable
Tree vote count for outcome variable
Final prediction

Figure . Random forest (RF) procedure flow chart representation.

2.6. J48 (Decision Tree Algorithm)

J48 or decision tree algorithm allows to calculate the feature behavior of different test groups. With J48, it is easy to understand the explanatory distribution of instances. This can help in identifying missing attributes and therefore works as a precision tool in case of over fitting was occurred [25]. The major challenge associated with the decision trees is the identification of the root node attribute. This attribute selection can be done in two methods: information gain (Equation 5) and Gini Index (Equation 6).

$$\text{Information gain written as Gain }(X, A) = \text{Entropy }(X) - \textstyle\sum_x \text{Values}(X) \frac{|Xx|}{|X|} * \text{ Entropy }(Xx) \tag{5}$$

Here, X: Set of instances, A: attribute, Xx: a subset of X with A = X, and value (A): set of total possible values of A.

Gini index (GI) is a parameter that helps to calculate how often randomly selected instances could be incorrectly classified.

$$\text{GI} = 1 - \sum_{a} z_a^2 \tag{6}$$

2.7. Performance Measures

Model performance was decided on the basis of accuracy, precision, recall, and F-scores. The performance measures with formulation and definitions are provided in Table 3.

Table 3. 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	$\frac{TP + TN}{TP + TN + FP + FN}$
PRECISION (P)	Rate of correct predictions	$\frac{TP}{TP + FP}$
RECALL (R)	True positive rate	$\frac{TP}{TP + FN}$
F-Measure	Used to measure the accuracy of the experiment	$2 * (\frac{P * R}{P + R})$

3. Results

Due to the issues raised with the model over fitting, exclusion of over-sampling and under-sampling PIDD data-sets were done during experiments.

3.1. Pruned Decision Tree

The J48 model classifier was exposed with the remained dataset (after removal of missing instances) to generate a pruned decision tree. The output pruned decision tree with plasma value as a central node is represented in Figure 5. It is obvious that plasma glucose concentration has the highest information gain, which could be considered as the highest risk factor for diabetes. Other risk factors such as multiple pregnancies, release of high levels of insulin, and lineage function also increased the chances of having diabetes. Generally, pregnant women who do not take much physical exercise have higher chances of gaining weight, which in turn increases the likelihood of having Type 2 diabetes.

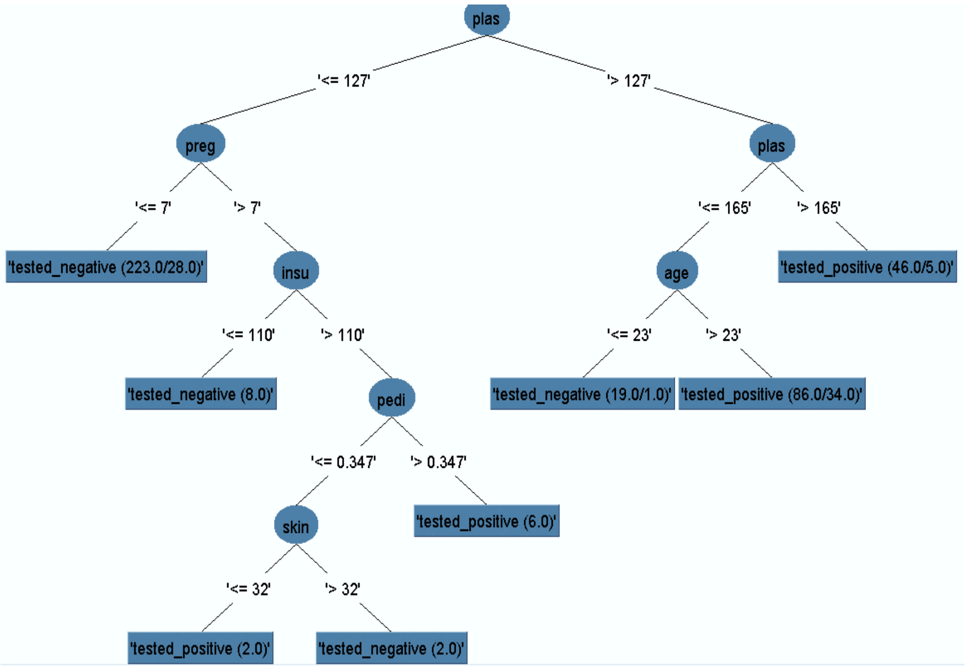

Figure 5. Pruned decision tree.

3.2. Confusion Matrix

The confusion matrix was used to describe the performance of various model classifiers [26]. The simulation was conducted with the four ML classifiers to analyze the accuracy of the prediction of the test class (Table 4).

Table 4. Confusion matrix of different classifier models.

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

3.3. Model Classification

We conducted the experiments with four ML classifiers to diagnose whether the patient was diabetic or non-diabetic. Table 5 shows the hyper parameters of four classifiers trained to classify diabetes of female patients. Performance measures validated all the models, exposed to different cross-validations to conduct model optimization techniques. The performance of four models chosen was depending on accuracy, recall, precision, AUC (area under the curve), and F-scores as shown in Table 6.

Table 5. Hyperparameters of different classifiers (here C: pruning confidence and 'R'–R squared value).

N	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

Table 6. Performance measures of different model classifiers (where k = 5, 10, 15&20).

K	Classifier	Accuracy	Precision	Recall	F-Score	AUC
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
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
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
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

Diabetes diagnosis at an early stage will give patients the opportunity to treat the disease and change their lifestyle in time to achieve positive results. In the present study, we propose an optimized machine learning algorithm for classifying and diagnosing Pima diabetic patients.

The majority of the results produced almost identical accuracy values. Hence, for assigning rankings for each model, Receiver Optimistic Curve (ROC) rates were used. ROC is a visualizing tool of the performance of the binary classifier. It is generated by plotting a false positive rate on the *X*-axis against the true positive rate on the *Y*-axis in order to decide the correct threshold value [27]. The AUC is the rate of accurate model classification and typically ranges between 0.5 and 1.0. If AUC is near to 1, the model performs correct classification of instances, and results in good optimization [28]. Four different machine learning algorithms were employed for various k values to predict whether a patient was diabetic or non-diabetic. Dataset was split into 'k' subsets to perform training and testing (in 'k' times). All preliminary analysis was carried out with the help of WEKA studio.

To avoid over fitting and under fitting issues, tenfold cross-validation was considered. The highest accuracy was achieved when the trained data had been exposed to k = 10. From Table 6, it was found that the LR model with the highest accuracy of 0.77, and NB, J48, and RF had an accuracy of 0.76, 0.73, and 0.74 respectively. In addition, recall (sensitivity) defines the rate of correctly predicted diabetic patients. For LR, it was found to be 0.77, and for RF, J48, and NB, it was recorded as 0.74, 0.73, and 0.76. The precision of NB was 0.75 that of J48 was 0.73, RF was 0.74, and LR was 0.76. F scores of J48, NB, RF, and LR were 0.73, 0.76, 0.74, and 0.76 respectively. In addition, we calculated the AUC to measure the performance of the four models. The AUC of J48, NB, RF, and LR was generated as 0.75, 0.81, 0.81, and 0.83.

These results clearly show that the four classifiers had similar prediction accuracy with small differences and margins of error. However, LR was the most accurate and J48 was the least accurate. Ultimately, LR, NB, and RF were deemed to be the three best models for predicting whether a patient is diabetic or not. Furthermore, for K = (5, 10, and 20), the NB parameters for accuracy, precision, recall, and f-scores were higher than those of RF. However, for K = 15, the RF precision and F-scores were higher than those of NB. Accuracy was not only the parameter, which can be used in assessing model optimization. The main limitation in using accuracy as the key performance metric is that it does not work well in datasets. This can generate class imbalances. The PIDD (Table 1) contains 500 women who tested negative for diabetes, and 268 women who tested positive for diabetes, and thus the imbalance ratio is 1.87. Hence, along with accuracy, it is also important to consider the AUC values (Figure 6). The AUC values of NB (Figure 6.1) and LR (Figure 6.2) were 0.81 and 0.83, respectively, and for RF (Figure 6.3), it was 0.82 and 0.81. However, J48 produced a lower AUC value (0.72) than others (Figure 6.4). When each classifier is ranked according to performance values, once seems that an optimized model is LR > RF > NB > J48.

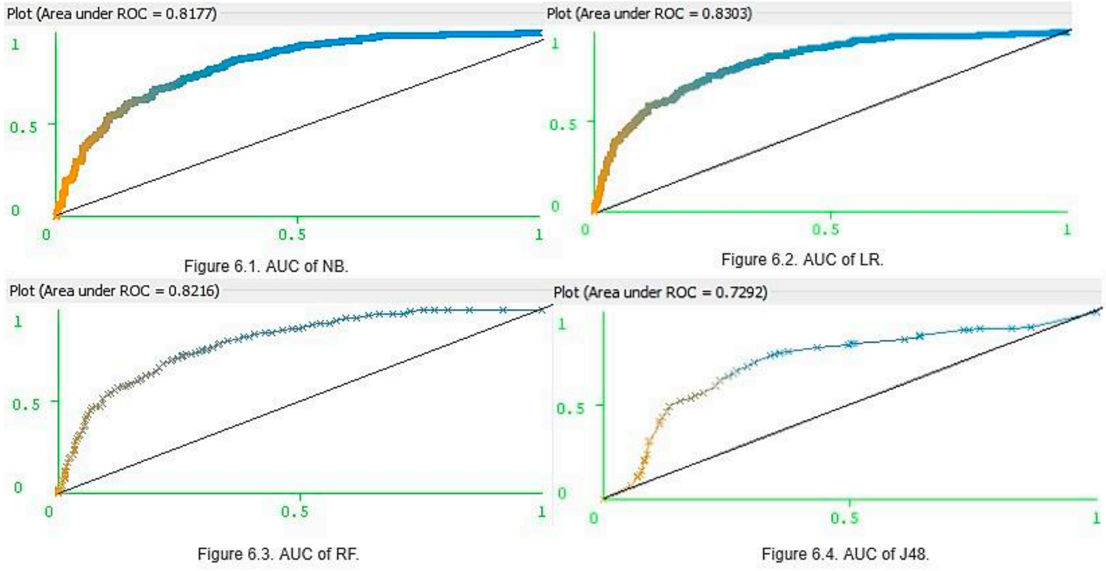

Figure 6.1. AUC of NB.

Figure 6.2. AUC of LR.

Figure 6.3. AUC of RF.

Figure 6.4. AUC of J48.

Figure 6. Area under the curve (AUC) of four different classifiers.

5. Conclusions

Diabetes is one of the most critical chronic diseases today, and early diagnosis can help greatly in improving a patient's chances of managing it well. The latest developments in machine intelligence can be exploited to improve our understanding of the factors causing the onset of this disease. We developed four binary classifier models: NB, J48, LR, and RF, and each model was analyzed using different CV methods (subject to different 'k' values). Performance assessment was conducted with the parameters of accuracy, precision, recall, F-scores, and AUC. Preliminary outcomes suggested that all models investigated achieved good results, with the LR model showing the greatest accuracy (0.77), and the J48 the relatively low accuracy compared to the others.

Ranking conducted by considering not only accuracy but also other parameters, and indicated that LR, NB, RF are the three best models for predicting whether a patient is diabetic or not.

The main limitation of this stdy is that only the conventional ML classifiers were considered. Since the results provide an improvement on existing methods for predicting diabetes, it would be worthwhile in future studies to explore these models in unsupervised machine learning and deep learning techniques as well.

Conflict of Interest: None

Acknowledgments: This study was supported by an institutional grant of Camerino University. Gopi B, Getugamo S, Nalini Ch Ph.D. bursaries were supported by the University of Camerino. The English language editing by Ms. Sheilla Beatty is gratefully acknowledged.

Author Contributions: We certify that the manuscript is not under review by any other journal. All authors have read and validated the final copy of this manuscript. GB*: Design and perform the experiments. Analyze the methods, results, and wrote the manuscript. CN& GS: Contribute to the literature review, SKT & FA: Conclusion and final manuscript revision.

References

1. Seshasai, S. R. K. et al. Diabetes mellitus, fasting glucose, and risk of cause-specific death. N. Engl. J. Med. (2011). doi:10.1056/NEJMoa1008862
2. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. The Lancet (2017). doi:10.1016/S0140-6736(17)30058-2
3. Kaur, H. & Kumari, V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Informatics (2019). doi:10.1016/j.aci.2018.12.004
4. Baştanlar, Y. & Özuysal, M. Introduction to machine learning. Methods Mol. Biol. (2014). doi:10.1007/978-1-62703-748-8_7
5. Battineni, G., Chintalapudi, N. & Amenta, F. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). Informatics Med. Unlocked (2019). doi:10.1016/j.imu.2019.100200
6. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. New England Journal of Medicine (2019). doi:10.1056/NEJMra1814259
7. Methods, D. P. Data Preprocessing Techniques for Data Mining. Science (80-120). (2011).
8. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. (2010).
9. Mathotaarachchi, S. et al. Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiol. Aging (2017). doi:10.1016/j.neurobiolaging.2017.06.027
10. Parmar, C. et al. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front. Oncol. (2015). doi:10.3389/fonc.2015.00272
11. Nirala, N., Periyasamy, R., Singh, B. K. & Kumar, A. Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine. Biocybern. Biomed. Eng. (2019). doi:10.1016/j.bbe.2018.09.007
12. Giger, M. L. Machine Learning in Medical Imaging. J. Am. Coll. Radiol. (2018). doi:10.1016/j.jacr.2017.12.028
13. Forouhi, N. G. & Wareham, N. J. Epidemiology of diabetes. Medicine (United Kingdom) (2019). doi:10.1016/j.mpmed.2018.10.004
14. Forouhi, N. G., Misra, A., Mohan, V., Taylor, R. & Yancy, W. Dietary and nutritional approaches for prevention and management of type 2 diabetes. BMJ (2018). doi:10.1136/bmj.k2234
15. Barakat, N., Bradley, A. P. & Barakat, M. N. H. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans. Inf. Technol. Biomed. (2010). doi:10.1109/TITB.2009.2039485
16. Zou, Q. et al. Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. **9**, 1–10 (2018).
17. Sisodia, D. & Sisodia, D. S. Prediction of Diabetes using Classification Algorithms. in Procedia Computer Science (2018). doi:10.1016/j.procs.2018.05.122
18. Wei, S., Zhao, X. & Miao, C. A comprehensive exploration to the machine learning techniques for diabetes identification. in IEEE World Forum on Internet of Things, WF-IoT 2018 - Proceedings (2018). doi:10.1109/WF-IoT.2018.8355130

19. Frank, E., Hall, M. A. & Witten, I. H. The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Fourth Edition (2016). doi:10.1016/B978-0-12-804291-5.00024-6
20. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. (2010).
21. Bergmeir, C. & Benítez, J. M. On the use of cross-validation for time series predictor evaluation. Inf. Sci. (Ny). (2012). doi:10.1016/j.ins.2011.12.028
22. Patil, T. R. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. Int. J. Comput. Sci. Appl. ISSN 0974-1011 (2013).
23. Schein, A. I. & Ungar, L. H. Active learning for logistic regression: An evaluation. Mach. Learn. (2007). doi:10.1007/s10994-007-5019-5
24. Menze, B. H. 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). doi:10.1186/1471-2105-10-213
25. Tsang, S., Kao, B., Yip, K. Y., Ho, W. S. & Lee, S. D. Decision trees for uncertain data. IEEE Trans. Knowl. Data Eng. (2011). doi:10.1109/TKDE.2009.175
26. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. (2009). doi:10.1016/j.ipm.2009.03.002
27. Lingenfelter, D. J., Fessler, J. A., Scott, C. D. & He, Z. Predicting ROC curves for source detection under model mismatch. in IEEE Nuclear Science Symposium Conference Record (2010). doi:10.1109/NSSMIC.2010.5873935
28. Huang, J. & Ling, C. X. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. (2005). doi:10.1109/TKDE.2005.50