International Journal on Emerging Technologies **11**(3): 659-665(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation

Puneet Misra[1] and Arun Singh Yadav[2]
[1]Assistant Professor, Department of Computer Science, University of Lucknow, Lucknow (Uttar Pradesh), India.
[2]Research Scholar, Department of Computer Science, University of Lucknow, Lucknow (Uttar Pradesh), India.

(Corresponding author: Puneet Misra)
(Received 27 March 2020, Revised 06 May 2020, Accepted 08 May 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In Machine Learning (ML) community, researchers are proposing complex models for real-life problems to achieve higher accuracy, which requires high computing and other resources. Fields like computer vision and NLP have given rise to deep learning with complex and high computational models setting the trend to apply them in almost all the fields While they help where we have an abundance of data and complex relationships, simpler models still can do wonders and on their day can challenge these behemoths. Here feature selection plays an important role and drastically improves the model accuracy. We have proposed the Recursive Feature Elimination with Cross-Validation (RFECV) approach for Type-II diabetes prediction to improve the classification accuracy. The major challenge with this approach is to deal with overfitting issue and improve the accuracy without unnecessary record deletion. We have applied other pre-processing methods and then have applied five different classical ML algorithms Logistic regression, Artificial Neural Networks, Naïve Bayes, Support Vector Machine, and Decision Tree (DT) to predict diabetes onset. LR provided the best accuracy (84%), and the rest of the models remains very close to each other.

Keywords: Machine learning, Disease prediction, classification, Preprocessing, feature selection, Recursive Feature Elimination (RFE), Cross-Validation (CV), Logistic Regression (LR), Artificial Neural Networks (ANN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT).

I. INTRODUCTION

Intelligent learning for prediction and forecasting is the topic that is under consideration in today's promising research related to Artificial Intelligence (AI). Learning is a critical requirement for any intelligent behaviour. Researchers have agreed that without learning, there is no intelligence. Therefore, machine learning has become a rapidly developing subfield of AI research. These intelligent algorithms were from the very beginning designed and used to analyze medical, clinical information [1, 2]. Machine learning algorithms analyze the historical data and extract the useful and do the diagnosis and predictions by learning from the patterns [3, 4]. Challenges with medical data are that it is non-linear, heterogeneous, and noisy [5]. So that information needs to be pre-processed to get the better result. Diabetes is a severe health problem in which the amount of sugar content cannot be regulated. It causes by the insulin imbalance in the body wherein the case of type I diabetes, the human body does not produce enough insulin, and in type II, the body becomes insulin resistance. Thus, the early and timely diagnosis of diabetes may prevent serious complications. The various machine learning-based system has been developed in recent years to predict diabetes [6, 7] still scientists and medical experts evolving new and intelligent algorithms and proved that machine learning algorithms [8, 9] performed better in disease diagnosing. The capability to work on extensive, heterogeneous data taken from different sources and keep improving the model performance by adding the background details to make it a more powerful tool [10]. The only objective of these developed systems is to improve the accuracy that leads to the correct prediction of the disease.

In previous studies, it has been seen that the author has applied various feature selection techniques but has trimmed the dataset with inconsistencies. Another drawback is that they simply applied a complex algorithm on every problem, whether it is simple or complex as a black box approach.

This study proposes LR-RFECV method to improve the classification accuracy of the model and do a comparative study of classical supervised ML algorithms. So, the proposed method does not delete the records but focuses on each feature and record its importance with predictor variable. This study also establishes the fact that simpler models works better than complex if tuned effectively. We will investigate their logic, assumptions, feature selection, pre-processing impact, etc. and will show that the less complex algorithms can do better on less complex problems.

The article organized as following sections, section I provides a brief introduction about the work, section II contains the related work, sections III focuses on the adopted methodology for Type II diabetes prediction and section IV includes the comparison and discussion on the produced results by the various classifiers.

II. REVIEW OF LITERATURE

Machine learning is the problem of induction, where general rules are learned from domain-specific observed data. It is not feasible to know what representation or what algorithm is best on the given problem beforehand. Without knowing the problem so

well, a user probably don't need machine learning, to begin with. The machine learning model allows a section of pre-processing, which removes irrelevant information from the data sources. The removal of unwanted data must be done very carefully by understanding the nature of data and the correlation of different features. Because the highly correlated features have adversely affected the performance of various ML classifiers. ML provides a unique way to deal with multicollinearity by using the feature selection algorithms which select the subset of relevant features from the original one based on some specific criteria. In general, FS methods can be categorized as filter, wrapper and embedded methods [11]. The filter method uses the criterion functions which did not belong to the classifier while the other two select the features with the learning mechanism of the classifiers. Even the less complex ML algorithms like the logistic regression method have used wrapper method of FS for improving the classification accuracy. The LR predicts the value of dependent features using prior probability [10]. The outliers undoubtedly impact prediction accuracy. Davis and Offord (2013) [12] used distance-based outlier detection as a pre-processing method and proposed a modified prediction model for diabetes type II prediction. The model achieved 79% of accuracy by using the sigmoid function, but after applying the Neuro based weight activation function to calculate bipolar sigmoid, the accuracy reached 90.4% [13]. However, the above-discussed paper was unable to show how pre-processing impacts the prediction accuracy and which is shown in this article [14], which predicts the 30-day readmissions risk for diabetes patients by applying different pre-processing techniques. A very slight improvement can be seen in the Naïve Bayes model logistic regression and decision tree [14]. The problem of the highly skewed dataset can be overcome using subsampling, but the class imbalance problem cannot produce a good prediction model. Barakat *et al.,* (2010) proposed a hybrid diabetes prediction model using the SVM classifier. The author has used K-means clustering algorithms for the pre-processing scheme to handle the class imbalance problem [15]. A total of five clusters are derived from the dataset and every cluster positive samples are taken based on Euclidian distance. Here the SVM provides promising results for diabetes prediction with 94% accuracy. Jarullah (2011) has used the J48 decision tree classifier on the modified dataset (pre-processed data). After applying the unsupervised k-means clustering for class imbalance problem and numerical discretization to make small groups of each attribute, the author achieved 78.17 % of accuracy [16]. But the decision tree can do better and Chen *et al.,* (2017) [17] has used k-means and 10-fold cross-validation technique for data pre-processing. The author significantly improved the performance of the decision tree model. With this dataset, the author achieved 90.04% accuracy on the PIDD dataset. The outlier problem may produce the wrong result. Ramezani *et al.*, (2018) used multiple imputation methods for missing value treatment and OT for dimensionality reduction [18]. This modified dataset applied to the hybrid model LANFIS (Logistic Adaptive Network-based Fuzzy Inference System). This model has achieved 88.05% accuracy. Sometimes the uncorrelated variables reduce the performance of any learning model, so finding uncorrelated attributes means the principal components. Kanchan and Kishor (2016) used the PCA as a pre-processing scheme [8]. The modified dataset applied to classifiers, where SVM outperform after applying PCA. One of the closest works can be seen where the author has used the PCA and some other unsupervised ml methods for pre-processing. This pre-processed dataset then applied on ANN classifier, which predicts diabetes with 92.28% accuracy [19]. Model selection for the problem is the biggest challenge where even the less complicated models can make a better prediction but here, the quality of data plays a significant role. Wu *et al.*, (2018) has done excellent work on data using a feature selection approach with correlation check and k-means clustering [20]. They prepared the data so well that even the less complicated models like logistic regression classified the diabetic positive and negative patient with 95.42 % accuracy. Naïve Bayes always worked better for imbalanced and missing data [21]. It fairly achieved the 76.3% accuracy after applying k-means and weka tool filtering approach. Lydia *et al.,* (2019) have used the traditional way of feature selection and when it applied on three different disease datasets(diabetes, cancer and thyroid) Naïve Bayes has given 75 % accuracy with diabetes dataset and SVM outperform in rest of the two [22].

III. EXPERIMENTAL SETUP

In this study, we have used the famous Pima Indian Diabetes Dataset (PIDD) [23]. Pima Indian is a group of Native Americans living in Southern Arizona. Due to some genetic issues, they take a poor diet of carbohydrates. However, in recent years they moved towards processed food rather than traditional agriculture food with minimum physical activity. This sudden change in habit and food makes them the highest prevalence of type-II diabetes, which makes them a reason for the research. This database was taken from the UCI machine learning library [24]. PIDD is a benchmark for comparing methods and widely adopted free datasets for research purposes [25] in the machine learning community.

The experimental setup for this study is divided into two stages. The first stage deals with data-preprocessing(A) methods as we have seen in the literature review and previous study [26] that the data pre-processing has drastically improved the results. Stage A gives the final predictors which are used by stage B for training, testing, and predictions. All the experiments have done in this study on the jupyter notebook [27] using the python programming language. Here Ipython compiler is used to run python programs.

The methodology includes the data collection and analyzing the nature, the pre-processing methods, feature selection, and predictions. The model proposed for this study are as follows (Fig. 1):

A. Data Preprocessing

The PIDD contains 768 records of pregnant females with eight characteristics and one more column for the outcome.

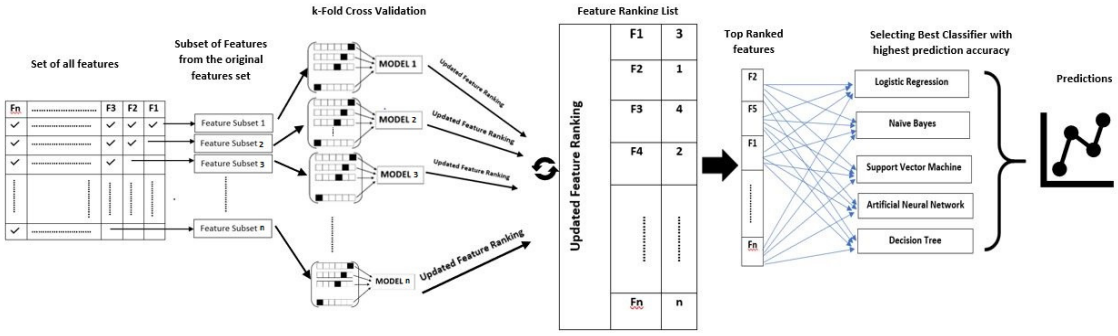


Fig. 1. Proposed model.

Each attribute is assigned the numeric value. In the dataset 65.10% (500 females) are non-diabetic (represented with value 0) and 34.90% (268 females) have diabetes (represented with value 1). We have used the following attributes of PIDD dataset (Table 1):

Table 1: The PIMA Indian Diabetes Dataset with a description.

S. No.	Parameter	Description	Data Type
1.	PREGNANT	Number of time women get pregnant	Numeric
2.	PGLUCOSE	Plasma glucose concentration measured using a 2-hour oral glucose tolerance test in mm Hg	Numeric
3.	DBP	Diastolic blood pressure	Numeric
4.	INSULIN	Two-hour serum insulin in muU/ml	Numeric
5.	TSFT	Triceps skinfold thickness in mm	Numeric
6.	BMI	Body mass index in mm2	Numeric
7.	DPF	Diabetes pedigree function	Numeric
8.	AGE	Age of the patient	Numeric
9.	OUTCOME	Patient with diabetes onset within five years (0 or 1)	Nominal

Missing value Treatment: The initial investigation of the dataset suggests that it is a supervised classification problem. The PIDD contains several inconsistencies in it as the metadata shows no missing values but Table 2 exhibits biologically implausible zero values.

Table 2: Occurrences of zero in different variables.

S. No.	Variable	No. of Zero
1.	PREGNANT	111
2.	PGLUCOSE	5
3.	DBP	35
4.	TSFT	227
5.	INSULIN	374
6.	BMI	11
7.	DPF	0
8.	AGE	0

This situation suggests that metadata is incorrect and must be treated as missing values. Some of the previously published studies have overlooked this and directly used them as recorded. However, this was a serious concern because INSULIN variable has more than 40% values are zero. After that, researchers start treating them as missing data and have published several studies. The occurrences of zero value in different variables are as follows:

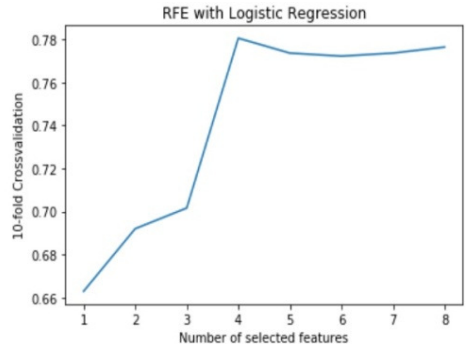
However, we cannot be sure in some cases that the presence of zero should be treated as missing or not. In the case of variable PREGNANT (number of times a woman gets pregnant), it can be zero times or more than one both cases can be considered but treat it as non-missing is more relevant than a missing instance. Missing data can severely distort the correlation between the variables. In the case of BMI and TFST, both variables used to measure obesity and must be highly correlated, but the computed correlation coefficient recorded 0.393, which is a weak positive correlation. After removing the record of zero instances of TFST yields correlation coefficient 0.632(highly positive). Instead of removing the missing instances, we have calculated the feature importance on the sample with no missing values using Recursive Feature Elimination (RFE) [28]. To get more confidence in feature selection k-fold cross-validation with Stratified k-fold is used.

Recursive Feature Elimination with Cross-Validation (RFECV): It is a wrapper method of feature selection. It removes the redundant and weak feature whose deletion least effects the training error and keeps the independent and strong feature to improve the generalization performance of the model. It uses the iterative procedure for feature ranking which is an instance of backward feature elimination. This technique first builds the model on the entire set of features and ranked the feature according to its importance. After that, it removes the least important feature and rebuilds the model again and recalculate the feature importance [29]. Let T is a sequence number to store the feature ranking. At each iterative process of backward feature elimination, the Ti stores the top-ranked features on which the model refit and performance is accessed. The value of Ti with the best performance is computed and top-performing features fit with the final model.

Algorithm 1: Recursive Feature Elimination with Cross-validation	
1.1	Train the LR model using all features with 10-fold cross-validation
1.2	Compute the model performance
1.3	Calculate the Feature importance or

	ranking
1.4	For each subset Ti, i=0,1,2,3,……., n do
1.5	Keep the Ti most important features
1.6	Train/Test model on Ti features
1.7	Recalculate model performance
1.8	Recalculate the importance of ranking of each feature
1.9	End
1.10	Calculate the performance over Ti
1.11	Determine the optimal number of features
1.12	Use the model with the selected optimal features

The top-performing features are based on Recursive Feature Elimination with Cross-Validation (RFECV) results are PREGNANT, PGLUCOSE, BMI, and DPF (Fig. 1). The selected features have few missing values which have replaced by the mean. After selecting the above features, we have used the complete dataset for the experiment. Scaling is performed on features at unit variance for efficient learning. The pre-processed dataset is then used in Stage B for the predictions.




```
The most suitable features for prediction:
['Pregnancies', 'Glucose', 'BMI', 'DiabetesPedigreeFunction']
```

Fig. 2. The most optimal features from the set of features given by RFECV approach for the type II diabetes.

B. Classification

Classification is the task of assigning the new observation to the class to which they most likely belong, i.e. close to the accuracy, based on the classification model built from the labeled training data. E.g., A good classifier can predict the condition of the patient in the future based on various symptoms and other parameters.

The classification can be binary and multilevel. When only two target classes are there in the problem, it is known as binary classification. For example, whether the patient has type-2 diabetes or not ? Nevertheless, in the multilevel classification, there must be more than one target class present in the problem statement. For example, a patient admitted in the ICU has a low, medium and high risk of mortality. The dataset taken for this study is a binary classification problem.

In the machine learning approach, the actual dataset is divided into two parts. The first part of the data (training data) is used to build the classification model by training it and the second part (test data) validates the model accuracy. Splitting of data must be done carefully else the information leakage can happen from test data. In this study, we have used the train_test_split() method of the Scikit-Learn library of python. Through this function, we divide the dataset into a different ratio. However, the 80/20 (train/test) rule is mostly used in the studies. We have used the following classification algorithms:

Logistic Regression (LR). It is a supervised machine learning algorithm borrowed from the traditional statistics which uses a Logistic function called sigmoid function g(z) that takes any value (independent variables) and predicts the discrete categories (dependent variables) between 0 and 1. But using the OvR technique, this model extended to multiclass classification. As it is borrowed from the linear regression, so the z value is similar to linear regression:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots$$
$$h(\theta) = g(z)$$
$$g(z) = \frac{1}{1 + e^{-z}}$$

The $h(\theta)$ means $p(y=1|x)$, i.e. the probability of predicted positive events. For example, the probability that the patient has type II diabetes, given features x. So, the inverse probability, not having disease $p(y=0 \mid x)=1-h(\theta)$. Logistic regression uses cross-entropy as a loss function due to non-linear sigmoid function at the end. The cost function will use two equations as given below:

$$J(\theta) = \frac{1}{m}\sum cost(y`, y)$$
$$cost(y`, y) = -log(1 - y`)\ if\ y = 0$$
$$cost(y`, y) = -log(y`)\ if\ y = 1$$

In this experiment, we have used GridSearch with k-fold cross-validation to find the best parameters for the LR. The parameters used with LR are given as follows:

Table 3: Parameters used in LR as returned by grid search.

Parameters	Values
C	1
Penalty	L1
Solver	newton-cg

Artificial Neural Network (ANN): In this study, we have used a Multilayer Perceptron (MLP) model of Artificial Neural Network. It is a supervised algorithm that learns from the labeled training set of the given data for the function f(.):Rm==>Ro. Here m represents the number of features given as an input vector whereas o denotes the number of features for the output vector. It learns the non-linear function approximation for regression or classification problem from the given independent variable X=x1,x2,x3….. and dependent variable Y. We have used MLP classifiers for our problem. The gradient descent approach used in ANN training. These gradients are calculated using back propagation which reduces cross-entropy loss function in classification. In this experiment Two-layer feed-forward back propagation neural network has been employed. For model tuning, we have used the Grid search method to find the optimal parameter setting of ANN. Parameters used in this experiment are given in table 4.

Table 4: Parameters used for ANN suggested by GridSearch

Parameters	Levels
Learning rate	Constant
Hidden layers	2
Activation	Relu
Maximum Iterations	500

Here four best features have been used to train the model thus the input layer comprises four neurons. The input layer of ANN has used four neurons where each represented as an optimal feature given by RFECV in stage A. One hidden layer was used with five hundred neurons set for the maximum iterations. Similarly, the result was obtained from another hidden layer with the constant initial learning rate 0.001 and activation function relu.

Naïve Bayes (NB): It is a probabilistic method that applies Bayes theorem. It calculates the probability of a given record belonging to a specific class. It assumes that given the class, features are statistically independent of each other. This assumption is called class conditional independence, which significantly simplifies the learning process. It is a generative method that generates examples from the assumptions and distributions. This prior knowledge then used by the model to predict the unseen data. It performs better on less training data despite naive assumptions. NB is always the best choice for quick and dirty implementation and considered to be the benchmark. In this experiment, we have used Gaussian Naive Bayes to predict likelihood. We have not used a grid search for NB because it has nothing to tune.

Support Vector Machine (SVM): Support vector classifier is also called the maximum margin classifier because it creates the maximum margin hyperplane. To achieve this the decision boundary defined to maximize the margin between the positive and negative classes. The window functions or kernels are responsible for converting the inputs into the required format. SVM have different types of kernels according to the problem like linear, non-linear, polynomial, radial basis function (RBF) and sigmoid. It returns the inner product of two points in a suitable feature space and thus can work well with a high dimensional dataset. In this experiment RBF, the most popular kernel is used. Gamma and C parameters are tuned to get the optimal values to achieve higher accuracy.

Table 5: Optimal parameter combination used in SVM using grid search.

Parameters	Levels
Kernal	Rbf
Gamma	0.05
Regularization (C)	12

Decision Tree (DT): This algorithm is inspired by the tree data structure where it constructs a hierarchical structure from the given training data. It divides the training data on the value of a feature. This model learns decision rules inferred from the features and predict the target class. In this experiment, we have used the CART (classification and regression tree) algorithm of decision tree because the training space has only numerical values. CART creates the binary tree using the features and threshold that yields the maximum information gain using the Gini index at each node. We have used the Decision Tree classifier from sklearn library that contains fourteen different parameters, but we tune only two parameters that are max_depth and min_samples_split to control the size and complexity of the tree. The optimal parameters used in the model are given in the table below:

Table 6: Best Parameters Given By Grid Search for DT.

Parameters	Values
Maximum depth of the model	3
Minimum samples to split	2

IV. EVALUATION MEASURES AND RESULT

Accuracy, sensitivity and specificity matrices are used in this experiment to evaluate the performance of predictions of the model. If the training space is balanced correctly, then the accuracy measure is enough to evaluate the model performance. However, in this experiment, the target variable is imbalanced, i.e. 34.9% are diabetic and 65.1% are non-diabetic patients that's why precision, recall, and F-score measures have used. To calculate all these measures confusion matrix is needed that are True Positive, False Positive, True Negative, False Negative. The formulation of the measures is given below in the table:

Table 7: Measures used for model evaluation.

Measures	Formulation
Precision(P)	TP/(TP+FP) & TN/(TN+FN)
Recall(R)	TP/(TP+FN) & TN/(TN+FP)
F1-Score	2*P*R/(P+R)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

All the tests were conducted on the discussed experimental setup and only the best results are taken of the discussed model evaluation matrices. The results of each prediction model are reported in table 8 and the comparative chart of the model performance is given in Fig. 3. The results produced by each model are satisfactory with the average accuracy of 80%, while the best is achieved around 84% by the LR. From the recorded values in Table 7, LR has identified diabetic patients with a high recall of 84% but at the same time, the model has classified the non-diabetic patients with 84% precision. The computed harmonic mean (F1-Score) of the precision and recall for LR is also 84%. The performance of ANN is very close to the best performing model and has achieved 81% accuracy. ANN is more complicated than LR, challenging to train, overfitting issues, difficult to optimize, and need a large numbers of training examples for generalization.

Table 8: Best result obtained from each model on used evaluation matrices.

Model	Precision	Recall	F1-Score	Accuracy
LR	0.84	0.84	0.84	0.837
ANN	0.81	0.81	0.81	0.817
NB	0.80	0.80	0.81	0.805
SVM	0.80	0.80	0.80	0.798
DT	0.80	0.81	0.80	0.805

In our case, ANN could perform better if we have more training examples and a balanced dataset. The Gaussian Naïve Bayes(NB) and the Decision Tree(DT) predicted with 80% accuracy, but DT predicted diabetic patients better than NB. The worst performance is given by SVM with 79% accuracy. The SVM performs worse with a small dataset this is because the data points near the support vectors (decision boundary) may not be a true representation of classification decision boundary and thus creates the false maximum margin hyperplane.

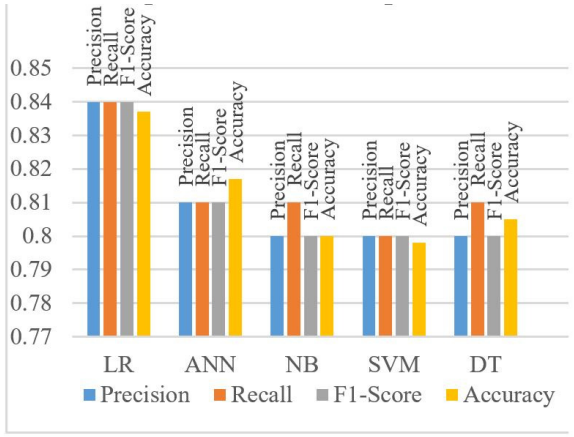


Fig. 3. Performance comparison of classifiers on different evaluation measures.

V. CONCLUSION

Classification problems with small sample sizes and many features have drawn the attention of ML practitioners. The feature selection is playing an important role to improve the classification accuracy of the ML models. In this experimental study, we have proposed LR-RFE with Cross-validation based feature selection method to classify Type II diabetic patients. As we have shown in the literature review that many complex ML models have accurately predicted Type II diabetic and non-diabetic patients with greater accuracy. However, we hypothesize that even the simplest ML model can do better than complex models if we properly examine the problem type and apply the suitable feature selection techniques. In previous studies, authors have trimmed the dataset to treat the inconsistencies but, in this experiment, we have taken a complete dataset. We have applied the RFECV method on the complete dataset and found that the features that contain the many missing values have minimal impact on the prediction accuracy and the top four features are giving the best result. To avoid the overfitting problem with RFE, we applied 10-fold stratified cross-fold validation. After including top-ranked features (PREGNANT, PGLUCOSE, BMI, and DPF), the pre-processed dataset then applied on different ML models, LR outperforms on all used model evaluation measures. While another complex model performs approximately the same on all measures. We have also observed that the feature selection method on the few dimensions (in our case 8 independent features) has contributed to improving the model accuracy and has helped to avoid the severe concerns like multicollinearity. We have also compared the results of our study with the previous one (as discussed in the literature review) and have found that our proposed method achieved greater accuracy.

Table 9: Comparison of our proposed method with existing work.

Methods	Accuracy Score	Reference	Year
J48 decision tree, discretization	78.1768%	Al Jarullah [16]	(2011)
Naive Bayes, SVM, Decision Tree, k-means on WEKA tool	76.30 %	Sisodia and Sisodia [21]	(2018)
LR, Naive Bayes, SVM, Decision Tree and k-nearest neighbor	75.12 %	Lydia *et al.*, [22]	(2019)
RFECV, LR, ANN, SVM, DT	84.00%	Our proposed method	2020

This study also tries to establish the fact that not every time highly complex models are needed for achieving more accuracy but even the less sophisticated models can give better accuracy. But this is not true in all respect and depends on the nature of data, its quality, volume, etc. It is also possible that complex models can give better results by going the deep dive in the problem set and its inconsistencies.

VI. FUTURE SCOPE

This study has been done on recursive feature elimination method and type II diabetes dataset has been used to justify the proposed approach. In future we can apply the same approach with different disease classification problems to make more generalized model.

Conflict of Interest. The author confirms that there are no known there are no conflicts of interest associated with this publication of the research article.

REFERENCES

[1]. Magoulas, G. D., & Prentza, A. (2001). Machine Learning in Medical Applications. *Machine Learning and Its Applications*, 300–307.

[2]. Jain, A., Ahirwar, M., & Pandey, R. (2019). A Review on Intutive Prediction Of Heart Disease Using Data Mining Techniques. *International Journal of Computer Sciences and Engineering*, *7*(7), 109–113.

[3]. Frandsen, A. J. (2016). Machine Learning for Disease Prediction. Thesis, Paper 5975.

[4]. Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89–109.

[5]. Menasalvas, E., & Gonzalo-Martin, C. (2016). *Challenges of Medical Text and Image Processing: Machine Learning Approaches*. https://doi.org/10.1007/978-3-319-50478-0_11.

[6]. Habibi, S., Ahmadi, M., & Alizadeh, S. (2015). Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. *Global Journal of*

Health Science, *7*(5), 304–310. https://doi.org/10.5539/gjhs.v7n5p304
[7]. Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ Open*, *3*(5), e002457. https://doi.org/10.1136/bmjopen-2012-002457
[8]. Kanchan, B. D., & Kishor, M. M. (2016). Study of machine learning algorithms for special disease prediction using principal of component analysis. In *2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC),* 5-10.
[9]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116.
[10]. Gambhir, S., Malik, S. K., & Kumar, Y. (2016). Role of soft computing approaches in healthcare domain: a mini review. *Journal of medical systems*, https://doi.org/10.1007/s10916-016-0651-x
[11]. Weston, J., Elisseeff, A., Schölkopf, B., & Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *Journal of machine learning research*, *3*, 1439-1461.
[12]. Davis, L. J., & Offord, K. P. (2013). Logistic regression:Modeling Conditional Probabilities. *Emerging Issues and Methods in Personality Assessment*, 273–283.
[13] Nirmala Devi, M., Balamurugan, A. A., & Reshma Kris, M. (2016). Developing a modified logistic regression model for diabetes mellitus and identifying the0 important factors of type II DM. *Indian Journal of Science and Technology*, *9*(4), 1–8.
[14]. Duggal, R., Shukla, S., Chandra, S., Shukla, B., & Khatri, S. K. (2016). Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India. *International Journal of Diabetes in Developing Countries*, *36*(4), 469–476.
[15]. Barakat, N. H., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *14*(4), 1114–1120.
[16]. Al Jarullah, A. A. (2011). Decision tree discovery for the diagnosis of type II diabetes. *2011 International Conference on Innovations in Information Technology*, 303–307.
[17]. Chen, W., Chen, S., Zhang, H., & Wu, T. (2017). A hybrid prediction model for type 2 diabetes using K-means and decision tree. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 386-390.
[18]. Ramezani, R., Maadi, M., & Khatami, S. M. (2018). A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. *Alexandria engineering journal*, *57*(3), 1883-1891.
[19]. Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., & Shahmoradi, L. (2017). Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset. *Fuzzy Information and Engineering*, *9*(3), 345–357.
[20]. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type II diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100–107.
[21]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*(Iccids), 1578–1585.
[22]. Lydia, E. L., Sharmil, N., Shankar, K., & Maseleno, A. (2019). Analyzing the performance of classification algorithms on diseases datasets. *International Journal on Emerging Technologies*, *10*(3), 224–230.
[23]. Rep, C. O. (2016). *HHS Public Access*. *4*(1), 92–98
[24]. PIMA INDIAN DIABETES DATASET. (1988). Retrieved April 15, 2018, from UCI Machine Learning Repository website: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
[25]. Idri, A., Benhar, H., Fernández-Alemán, J. L., & Kadi, I. (2018). A systematic map of medical data pre-processing in knowledge discovery. *Computer Methods and Programs in Biomedicine*, *162*, 69–85.
[26]. Misra, P., & Yadav, A. S. (2019). Impact of Preprocessing Methods on Healthcare Predictions. *Available at SSRN 3349586*, 144-150
[27]. Avila, D., Bussonnier, M., Corlay, S., Brian Granger, & Grout, J. (2014). Juypter Notebook with Ipython. Retrieved May 18, 2018, from http://jupyter.org/install
[28]. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). *Gene Selection for Cancer Classification using Support Vector Machines*. *Machine Learning 46*, 389–422.
[29]. Mathew, T. E. (2019). *A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis*. *10*(3), 55–63.

How to cite this article: Misra, P. and Yadav, A. S. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. *International Journal on Emerging Technologies*, *11*(3): 659–665.