Publications

Available Online

J. Sci. Res. **13** (3), 901-913 (2021)

JOURNAL OF SCIENTIFIC RESEARCH

www.banglajol.info/index.php/JSR

Consensus of Feature Selection Methods and Reduced Generalization Gap Model to Improve Diagnosis of Heart Disease

S. Gupta*, R. R. Sedamkar

Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India

Received 3 May 2021, accepted in final revised form 26 July 2021

Abstract

Enhancing the diagnostic ability of Machine Learning models for acceptable prediction in the healthcare community is still a concern. There are critical care disease datasets available online on which researchers have experimented with a different number of instances and features for similar disease prediction. Further, different Machine Learning (ML) models have different preprocessing requirements. Framingham heart disease data is multicollinear and has missing values. Thus, the proposed model aims to explore the differential preprocessing needs of ML models followed by feature selection in consensus with domain experts and feature extraction to resolve multicollinearity issues. Missing values have been imputed differently for each feature. The work also identifies optimal train set size by plotting a learning curve that provides a minimum generalization gap. When testing is done on this hyperparameter tuned model, performance is enhanced with respect to the F score weighted by support and stratification since the data is imbalanced. Experimental results demonstrate improvement in performance metrics, i.e., weighted F score, precision, recall, accuracy up to 3 %, and F1 score by 8 % for Logistic Regression Classifier with the proposed model. Further, the time required for hyperparameter tuning is reduced by 50% for tree-based models, particularly Classification and Regression Tree (CART).

Keywords: Learning curve; Validation curve; Weighted F score; Multicollinearity; Feature extraction; Machine learning.

doi: http://dx.doi.org/10.3329/jsr.v13i3.53290 J. Sci. Res. **13** (3), 901-913 (2021)

1. Introduction

India will soon become the heart disease capital of the world. It is estimated to account for 35.9 % of deaths by the year 2030 [1]. Many healthcare data is available online, which has been collected for experimentation with the collaboration of clinical and technological experts, as shown in Table 1 below. Several heart disease datasets have been tested to find the optimal one for experimentation. Cleveland and Statlog [2] have fewer instances, Hungarian has large amounts of missing values, Alizadehsani [3,4] has many features, and Arrhythmia does not have meaningful features. Lastly, Cardio Vascular Disease (CVD) dataset has too many instances, and since deep learning algorithms are beyond the scope of our work, it has not been experimented with.

* *Corresponding author*: shiwani.gupta@thakureducation.org

Table 1. Choice of dataset.

	No. of instances	No. of features	Missing Values	Imbalance
Statlog	270	13	N	N
Cleveland	303	13	Y	N
Hungarian	294	10	N	N
AlizadehSani	303	54	N	Y
Arrythmia	452	279	Y	N
Framingham	4240	15	Y	Y
CVD	70000	11	N	N

Framingham heart disease data [5] predicting ten-year Coronary Heart disease (CHD) with 4,240 instances and 15 features are preferred for experimentation due to the availability of more data and less percentage of missing values. Any disease dataset has been imbalanced since the number of instances in the healthy class will be more than the diseased class.

Bias is an error from erroneous assumptions in a learning algorithm, whereas variance is an error from sensitivity to small fluctuations in the training set. The major issue with ML models is the bias-variance trade-off. Thus, to avoid the underfit/overfit of the model onto data, it is required to assess whether the data is linear/nonlinear. Secondly, underfit happens with fewer data; that is why other datasets were not chosen for experimentation. On the contrary, complex models get highly trained on data and hence overfit. They yield low performance on the test set.

k Nearest Neighbor (kNN) algorithm has been applied for Missing Value Imputation (MVI) onto a complete heart disease data to validate that it provides meaningful imputation since the data distribution does not change post imputation [6]. Extensive review has been conducted with respect to the MVI, feature selection methods, supervised ML models, and several heart disease datasets [7], which have helped us design the proposed methodology.

Experimentation with oversampling, undersampling, cost-sensitive classification, and an ensemble of cost-sensitive Decision Tree (DT) on imbalanced data has been done in reference [8]. The performance has been evaluated using Mathews Correlation Coefficient (MCC). Synthetic Minority Oversampling Technique (SMOTE) has been used to balance data along with evaluation measures as a confusion matrix, Stratified k fold, accuracy, Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) along with Logistic Regression, Support Vector Machine (SVM) and Artificial Neural Network (ANN) models [9]. Cost-Sensitive Learning (CSL) takes prediction error into account during the training process to better predict performance.

A combination of feature selection and imputation techniques has been proposed and experimented with for medical datasets where careful selection methods are required [10]. Genetic Algorithm (GA), a wrapper-based technique, and information gain, a filter-based feature selection technique, work well for low dimensional data, and Decision Tree, an embedded feature selection technique, works well with high dimensional data. DT has been found to filter useful features as well. Combining feature selection and extraction has proven to give better accuracy, sensitivity, and specificity utilizing SVM Radial Basis

Function (RBF) kernel [11]. Backward elimination (wrapper) and Pearson correlation (filter) based feature selection methods have been used to select major predicting features [12]. The performance was evaluated utilizing Naïve Bayes (NB) classifier. Missing values were replaced by mean categorical data. Similarly, a comparison of oversampling and undersampling techniques demonstrated enhancement inaccuracy by 20%. Sequential Feature Selection (SFS) based Wrapper with Random Forest (RF) has returned considerably high accuracy compared to other feature selection and ML model combinations. Even Least Absolute Shrinkage and Selection Operator (LASSO), an embedded feature selection technique, has better accuracy than different feature selection and ML model combinations [13]. Mutual information (MI) for feature selection has been used to increase classification accuracy and reduce execution time, along with Leave One Out Cross Validation (LOOCV) for model assessment and hyperparameter tuning [14]. Models experimented with are SVM, Logistic Regression (LoR), kNN, NB, and DT. A heterogeneous hybrid feature selection method integrated with a balancing approach has been tested on well-known Coronary Artery Disease (CAD) datasets utilizing DT, RF, Gaussian NB (GNB), eXtreme Gradient Boost (XGB), kNN, and Binomial NB (BNB) classifiers [15]. Features having a maximum correlation with output were used for training and tested on SVM, Linear Discriminant Analysis (LDA), CART, NB, kNN, and RF classifiers. Accuracy and Kappa score with 95 % confidence level were utilized for performance evaluation, demonstrating better ensemble performance over basic ML models [16]. Genetic Algorithm (GA) with SVM and GA with ANN wrapper has been utilized for identifying both optimal feature subset and hyperparameter values onto Framingham heart disease data with increased sensitivity and F1 score [17]. Since the dataset was small, Cross-Validation (CV) prevented overfitting.

A 2-tier classifier ensemble of RF, Gradient Boosted Machine (GBM), and XGB has been evaluated on multiple heart disease datasets in terms of accuracy, F1, and AUC scores [18]. A 2-level stacking with 10 base models where the dataset was randomly shuffled and split for 10-fold CV reported mean and standard deviation.

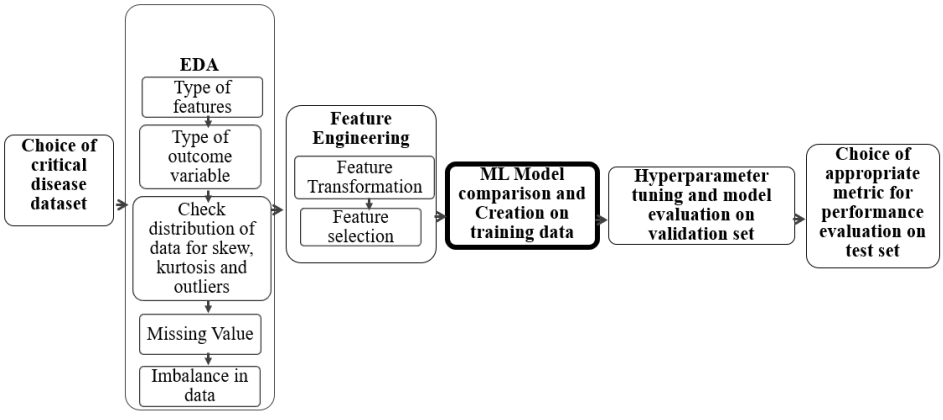


Fig. 1. Proposed methodology.

Fig. 1 above demonstrates the ideology in terms of choice of the appropriate dataset, followed by extensive Exploratory Data Analysis (EDA) in terms of the type of features, type of outcome variable, distribution of data in terms of skewness and kurtosis, check for outliers, missing values and imbalance in data. Further, feature engineering in terms of hybrid feature selection and extraction technique is proposed and experimented with.

The next step is to decide which ML model to train data on. Thus, comparing several models as Probabilistic Support Vector Classifier with RBF kernel, LoR with optimal threshold, BNB, kNN, CART, Extreme RF, and GB ML models has been made in terms of their preprocessing requirement.

A bagging algorithm as Extreme RF, i.e., weighted by support in comparison to RF, which is greedy, and GB to overcome the underfit of DT has been used. Adaptive Boosting (Adaboost) works best with weak learners; hence it is not chosen, and XGB works well with unstructured data; hence we do not choose it. The appropriate tuning of hyperparameters can improve the performance of each model. All heterogeneous models are sensitive to multicollinearity; hence the experimentation focused on multicollinearity for these 4 classifiers. All models require resampling except kNN. kNN and SVM RBF kernel require scaling. Only CART and boosting algorithms are robust to missing values. SVM, LoR, CART, and Bagging ML models are robust to the presence of outliers.

In none of the literature referred, the trade-off between bias and variance property of ML models has been explored. Among the available feature selection methods, if a single one is chosen for identifying the optimal feature subset, then the model would be biased. The hypothesis says, if the consensus of feature selection algorithms along with domain expertise is taken, then bias due to the different feature selection methods can be reduced.

Secondly, multicollinearity is a bane to data. Generally, researchers apply Principal Component Analysis (PCA) onto the entire dataset to reduce multicollinearity. Thus, the second hypothesis says that if feature extraction is performed only on multi collinear features with extremely high Variance Inflation Factor (VIF) than on the entire data, then performance can be enhanced in lesser time.

Imbalance in data can be treated either through an appropriate sampling technique that generates synthetic data to balance the classes or an appropriate ML model that deals with imbalance inherently. The hypothesis says, if an appropriate model evaluation technique as stratification along with CV is combined with weighted F score as the evaluation metric to handle imbalance, then performance could be better than an appropriate ML model or an appropriate sampling technique as SMOTE or Adaptive Synthetic (ADASYN) for Imbalanced Data.

Further, different datasets have a different number of instances. So, what should be the optimal size of training data? Hence, the last hypothesis states; if the train set size with minimum generalization gap is chosen, there is no need for more data to enhance the performance.

2. Experimental

2.1. *Algorithmic design*

Thus considering all the above aspects, the algorithm designed has the following steps:

(i) Perform data preprocessing
   (a) Check for skewness, kurtosis, and outliers: treat extreme outliers only since features are meaningful;
   (b) Data has missing values: treat each feature differently based on the number of missing values;

(ii) Perform Exploratory Data Analysis
   (a) Check multicollinearity: treat only features having very high VIF utilizing PCA;
   (b) Check Correlation: perform feature selection in consensus with domain expert;

(iii) Perform Model Building
   (a) Build all probabilistic and tree-based models with different preprocessing suited to it;
   (b) Plot validation curve to check whether the model is suffering from bias or variance: perform hyperparameter tuning to reduce bias and generalization gap;

(iv) Perform Model Evaluation
   (a) Compare model performance on train set size with minimum generalization gap in comparison to the static split of 70:30;
   (b) Since data is imbalanced, choose the appropriate performance metric as a weighted f score and model evaluation technique as stratified k fold CV.

2.2. *Data preparation*

Framingham heart disease data has many instances and has samples of both male and female genders in equal proportion. Even distribution of CHD in both genders is similar, which shows the appropriateness of chosen data. The categorical data distribution shows nonlinearity in data. Irrespective of the cohort taking Blood Pressure (BP) medication have had a stroke, hypertension, diabetes, is currently smoking, or the gender being male, he/she may/may not have the disease.

The features are meaningful, and the data is imbalanced. Skewness refers to lack of symmetry in features, and kurtosis refers to the heavily/lightly tailed data relative to the normal distribution. The acceptable range for skewness falls between -3 and +3, and kurtosis is appropriate between -10 to +10. This data is skewed for 7 features and has meaningful outliers except for age. Kurtosis exists in 4 features, but data has not been transformed because skewness and kurtosis provide meaningful information.

The proposed model compares all probabilistic and tree-based ML models which exhibit different properties. Probabilistic models are capable of returning a probability of an instance lying into a class than just the class label. Linear ML models specify that they can be represented as a linear combination of features. All chosen models are binary as the outcome variable of the chosen data is binary. A discriminative model models the decision

boundary between classes, whereas a generative model explicitly models the actual distribution of each class. In parametric models, the parameters do not change once the model is designed. Thus, all ML models can be implemented as probabilistic. kNN, CART, and RF are nonlinear models. All can be utilized for binary classification tasks. NB is generative; the rest all are discriminative models. LoR, NB, and RF are parametric, while others are nonparametric ML models.

2.3. *Missing value imputation*

Since 'education level' (2.5 % missing) is a subjective feature, it is dropped as it is not very handy to practice. Mean is the arithmetic average across the column. Median is the middle number in the column when data is arranged in either ascending/descending order, and mode is the value that occurs most often in the data. In the total cholesterol 'totchol' (1.2 % missing) feature, extreme values were dropped, then group median was performed. Only one instance had a missing value for Heart Rate' HR' (1 missing in 4240), so that was dropped. Group median performed for Body Mass Index 'BMI' (0.4 % missing) and glucose (9.2 % missing). Extreme instances for Cigarettes Per Day' cigsperday' (0.7 % missing) were dropped. Missing values in BP Medication' BPMeds' (1.2% missing) were imputed with mode. Since imputation is meaningful, the histogram doesn't change much post imputation.

2.4. *Imbalance*

Since data is imbalanced, stratification has been used. Stratification seeks to ensure that each fold is representative of all strata of the data. This is done to alleviate the bias of most classification algorithms that tend to weigh each instance equally. Thus, overrepresented classes get too much importance. Even mean weighted f score is reported to deal with the imbalance in data. The average is weighted by support, which is the number of samples per class. Precision is a metric that calculates the percentage of correct predictions for the positive class. In contrast, recall calculates the percentage of correct predictions for the positive class out of all positive predictions that could be made. There is a trade-off between precision and recall; hence f measure, a harmonic mean giving equal weightage to both is used.

2.5. *Feature selection and extraction*

Multicollinearity refers to the occurrence of high intercorrelations among two or more independent variables. Data being multicollinear could affect the performance of ML models; hence VIF has been computed, which is a measure of overall model variance to the variance of the model when it includes only a particular feature.

VIF exceeding 10 are signs of serious multicollinearity. Systolic BP ('sysBP') and Diabolic BP ('diaBP') are found highly correlated; hence feature extraction using PCA [19] is done for these, which captures 93 % variance in the data. Though Prevalent

Hypertension ('preHyp') was also correlated with 'sysBP' and 'diaBP', this was retained as advised by a medical practitioner. Similarly, 'currentSmoker' feature, which is highly correlated with 'cigsPerDay', has been dropped in consultation with a domain expert and with the consensus of multiple ML algorithms for feature selection as GANN, GASVM, Backward Elimination (Wrapper) [20], Chi^2 (Filter), logit, ExtraTree_importance (embedded), RF_importance (embedded), etc.

3. Results and Discussion

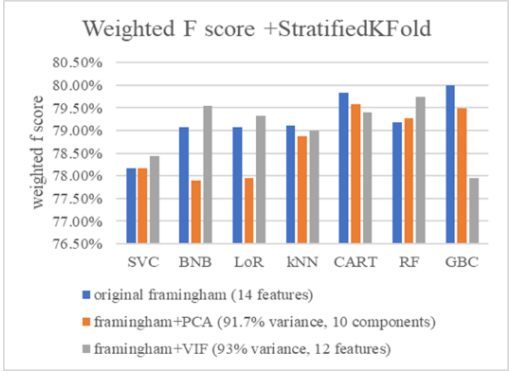
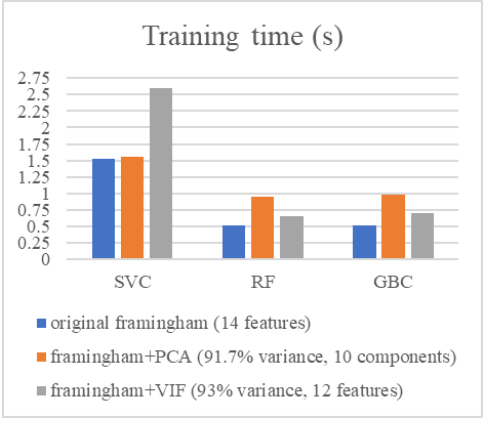


Fig. 2. Comparison of weighted F score.



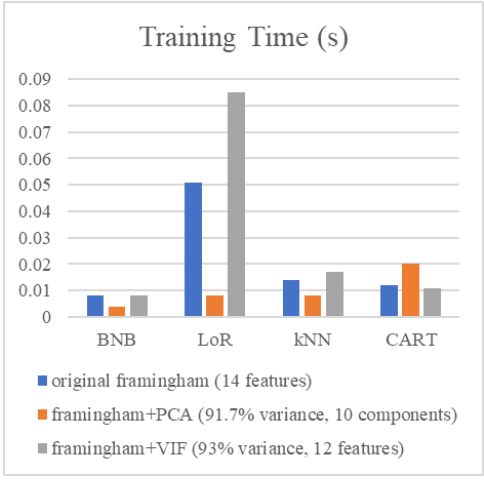


Fig. 3. Comparison of training time.

The models were tested after performing preprocessing of missing values and outliers in the manner described in sections 2.2 and 2.3 above. Performance evaluation was done on three datasets: original Framingham, one treated for multicollinearity (by applying PCA on entire data capturing 91 % variance with 10 components), and novel algorithm (by applying PCA on only 2 features with very high VIF capturing 93.7 % variance). The

performance metric used is a weighted F score for evaluation since data is imbalanced. Models have also been compared with respect to the training time. The results demonstrated in Fig. 2 above state that performance is enhanced for SVC, BNB, LoR, kNN, and RF. Initially, stratification was used for all models except kNN and GBC (insensitive to imbalance). Further experimentation with stratification improved the performance of kNN as well. Thus, the proposed methodology works well with heterogeneous classifiers (BNB, SVC, kNN, and LoR) to improve weighted F scores. For tree-based homogenous models, the proposed method reduces training time, as demonstrated in Fig. 3 above.

The hypothesis says that increasing the size of training examples does not guarantee enhanced performance. In order to test this, learning curves (Figs. 5 and 7) were plotted for base models, which were tree-based homogenous and probabilistic heterogenous as per Fig. 4 below.

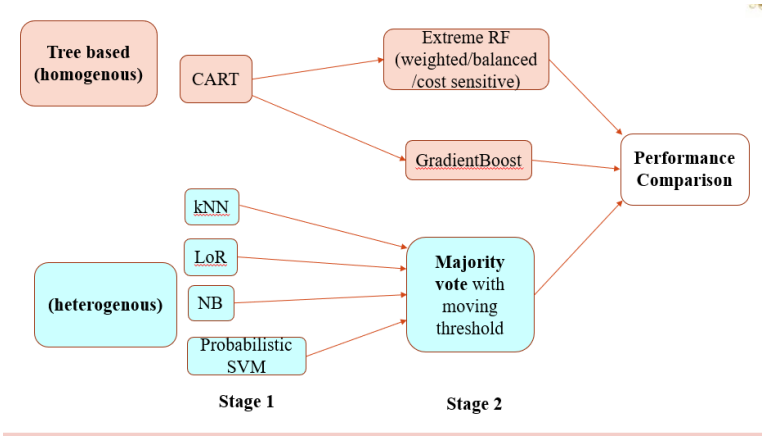


Fig. 4. Proposed general architecture.

All these models were highly biased particularly RF, since error did not improve with further training. This was apparent from the learning curves demonstrated in Fig. 5 below where performance in terms of weighted F score on training and validation set is shown with respect to the number of instances.

In order to enhance generalization, adding features was not feasible; hence hyperparameter tuning was performed utilizing validation curve with Randomized Search CV, since data had outliers. It sets up a grid of hyperparameter values and selects random combinations to train the model and score. The generalization gap was reduced by 3 % through hyperparameter tuning and further by 7 % through feature selection for the DT classifier. Hyperparameter tuning RF classifier reduced generalization gap as well. Feature selection reduced it further, but there was no improvement in the validation score. GB classifier has tree-specific and boosting hyperparameters. Initially, the boosting hyperparameter range was tested, which lowered the performance. Then tree-based

hyperparameters were tested sequentially, which reduced the generalization gap by around 4 % without reducing the validation score. The same is illustrated in Fig. 6 below.

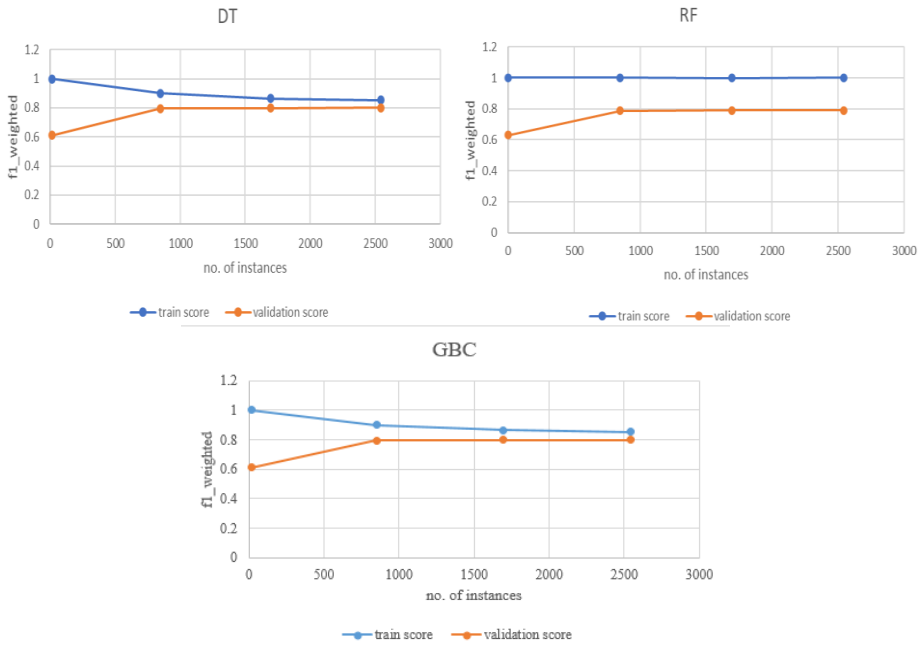


Fig. 5. Initial Learning curves demonstrating generalization gap for tree-based models.

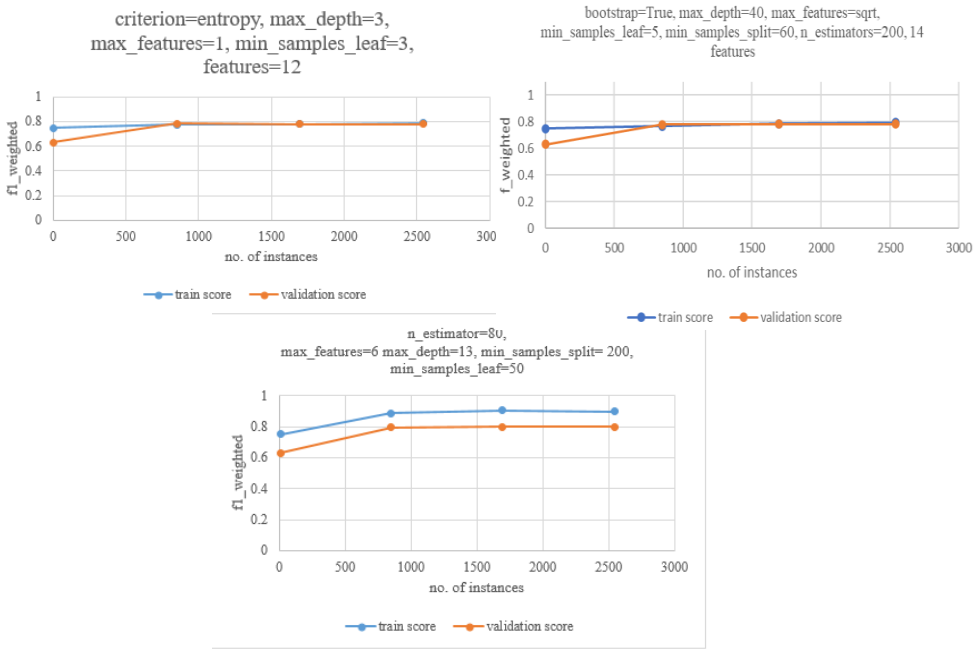


Fig. 6. Reduced Generalization gap for tree-based models post hyperparameter tuning.

For LoR, hyperparameter tuning did not improve performance. Scaling with RobustScaler for LASSO also did not reduce the generalization gap. For NB, removing multicollinearity improved performance on both train and validation set by 1 %. Moreover, NB does not have hyperparameters. Hence, the optimal threshold utilizing the Precision-Recall curve (PRcurve) has been experimented with. The generalization gap for kNN is reduced by identifying the optimal hyperparameters. Similarly, for SVM, hyperparameter tuning improved performance on both train and validation set by 1 %. The same is demonstrated in Fig. 8 below wrt Fig. 7 with initial Learning curves.

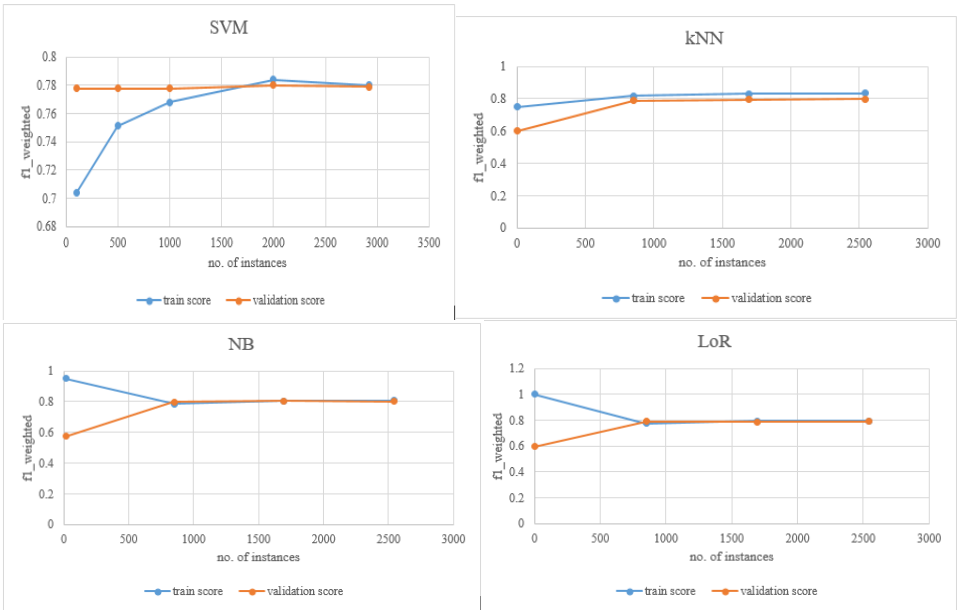


Fig. 7. Initial learning curves demonstrating a biased model.

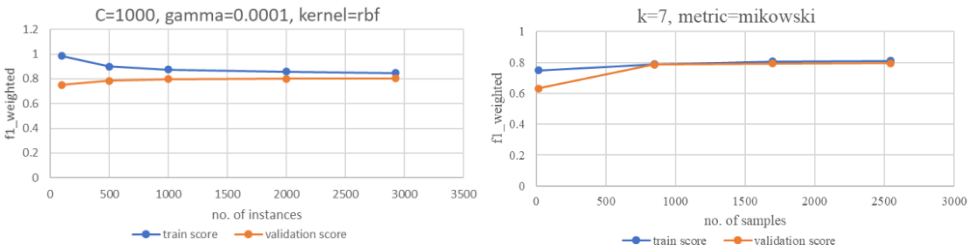


Fig. 8. Reduced Generalization gap for SVM and kNN post hyperparameter tuning.

Table 2 demonstrates the performance of various models and their optimal hyperparameters, comparing the performance of a novel approach with a traditional one and the time taken to hyper tune. Table 3 below shows the performance comparison with respect to several evaluation metrics, where LoR classifier shows improved performance with respect to weighted F-score, Precision, Recall, Accuracy, and F1score.

Table 2. Comparing the performance of tree-based classifiers with minimum generalization gap with respect to static split, also the time is taken to hyper tune.

Model	Hyperparameters	Optimal Train set size	Novel Approach	Traditional Approach	Time to hyper tune
kNearest Neighbor	k=7, metric=minkowski	600	79.4+/-0.023	78.7+/-0.010	-
Support Vector Classifier	C=100, γ=0.0001	400	77.9+/-0.002	78+/-0.003	8.2 s
Logistic Regression	C=2.07, penalty=L2	1500	73.1+/-0.032	70.5+/-0.027	38.1 s
Bernoulli Naïve Bayes	NA	1950	80.1+/-0.017	79.5+/-0.013	NA
CART	criterion=gini, max_depth=3, max_features=4, min_samples_leaf=4	750	78.6+/-0.019	78.2+/-0.007	7.3 s
Random Forest	n_estimators=200, min_samples_split=50, min_samples_leaf=10, max_features='auto', max_depth=25, bootstrap=True	1450	78.6+/-0.012	77.8+/-0.003	1.5 min
Gradient Boosting Classifier	learning_rate=0.1, n_estimators=80, max_depth=13, min_samples_split=200, max_features=6, min_samples_leaf=50, subsample=0.8	550	79.8+/-.025	79.8+/-.018	29.7 s

Table 3. Comparing the performance of probabilistic classifiers with static split and split with minimum generalization gap utilizing several performance metrics.

Methodology	train set size	Weighted f score	Precision	Recall	Accuracy	F1 score
Balanced Logistic Regression with optimal threshold=0.20	1500	73.1±3.2	28.6±4	69.7±8.7	68.9±3.8	45.6
	2958	70.5±2.7	25.7±3.1	66.5±7.8	65.7±3.2	37.1
Support Vector Classifier	400	78±1.2	NR	NR	84.9±0.9	NR
	2958	78±0.3	13.3±34	0.3±0.8	84.8±0.2	6
Bernoulli Naïve Bayes	1950	80.1±1.7	41.5±18	10.6±5.7	84.4±1.1	16.8
	2958	79.5±1.3	47.2±21.7	7.3±4.2	84.5±1.1	12.6
K Nearest Neighbor	1150	79.4±2.3	43.9±4.8	5.8±6.8	85±1.6	3.2
	2958	78.7±1	40.1±27.7	3.9±2.9	84.5±0.7	7.1

CV is a statistical method to evaluate ML models that result in lower bias estimates than other methods. The results of k-fold CV run are returned as mean with standard deviation (s.d.) as a measure of the variance of scores across folds. Since the data is imbalanced, stratification has been used.

4. Conclusion

The existing concepts have been ruled out through experimentation; for instance, kNN handles imbalance but still stratification improved performance. kNN and BNB are robust

to outliers; still, they gave a better performance on data not treated for outliers. The novel approach for multicollinearity treatment has reduced training time and increased performance for particular ML models. The hybrid feature selection and extraction method have provided better results in terms of weighted F score with SVC, BNB, kNN, LoR, and RF models compared to traditional feature selection methods. Suppose a different train set size is chosen for particular models based on the minimum generalization gap between the training and validation set (optimal hyperparameters). In that case, performance does not drop in weighted F score for all probabilistic and tree-based models. Further, Balanced Regularized LoR with optimal threshold has given better performance in terms of improved weighted F score, Precision, Recall, and Accuracy up to 3 % and F1score by 8 %. The time required for hyperparameter tuning is reduced by 50 % for tree-based models esp. CART.

Further to bring flexibility, probability prediction has been targeted. Since in the medical domain, the cost of False Positive (healthy predicted as diseased) is less than the cost of False Negative (CAD predicted as normal), probabilities can be interpreted by varying the thresholds. This has been done particularly for LoR using the PR curve. The future work lies in modifying the algorithm for optimizing the loss function of the LoR classifier [21,22] to enhance the performance further compared to complex ML models.

5. Acknowledgments

I am grateful to A. K. Rastogi, Dwarika Hospital, Muhamdi Khiri, UP, the domain expert consulted for feature selection. He also suggested that data need not be treated for outliers since they are meaningful. I am also thankful to A. Babbar, a data science intern who suggested not to treat data in terms of skewness and kurtosis as they are meaningful.

References

1. F. Babič, J. Olejár, Z. Vantová, and J. Paralič, ACSIS **11**, 155 (2017). https://doi.org/10.15439/2017F219
2. C. B. Gokulnath and S. P. Shantharajah, Cluster Comput. **22**, 14777 (2019). https://doi.org/10.1007/s10586-018-2416-4
3. M. A. Hogo, SN Appl. Sci. **2**, 1060 (2020). https://doi.org/10.1007/s42452-020-2858-1
4. M. G. Mohammad, S. Zendehboudi, and A. A. Mohsenipour. Comput. Methods Programs Biomed. **192**, 105400 (2020). https://doi.org/10.1016/j.cmpb.2020.105400
5. H. A. G. Alsayed and L. Syed, An Automatic Early Risk Classification of Hard Coronary Heart Disease using Framingham Scoring Model - *ICC '17: Proc. of the 2nd Int. Conf. on Internet of Things, Data and Cloud Computing* (2017). https://doi.org/10.1145/3018896.3036384
6. T. Razzaghi, O. Roderick, I. Safro, and N. Marko, PLoS ONE, **11**, ID e0155119 (2016). https://doi.org/10.1371/journal.pone.0155119
7. S. Gupta and R. R. Sedamkar, Apply Machine Learning for Healthcare to Enhance Performance and Identify Informative Features – *Proc. of IEEE INDIACom; 6th Int. Conf. on "Computing for Sustainable Global Development*, BVICAM (New Delhi, India, 2019).
8. W. Zheng, X. Zhu, Y. Zhu, and S. Zhang, Robust Feature Selection on Incomplete Data - *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence* (2018). https://doi.org/10.24963/ijcai.2018/443

9. I. C. Dipto, T. Islam, H. M. M. Rahman, and M. A. Rahman, J. Data Anal. Info. Process. **8**, 41 (2020). https://doi.org/10.4236/jdaip.2020.82003
10. C. H. Liu, C. F. Tsai, K. L. Sue, and M. W. Huang, Appl. Sci. **10**, 2344 (2020). https://doi.org/10.3390/app10217789
11. S. Shah, S. Muhammad, F. A. Shah, S. A. Hussain, and S. Batool, Comput. Electric. Eng. **84**, ID 106628 (2020). https://doi.org/10.1016/j.compeleceng.2020.106628
12. J. Wang, C. Liu, L. Li, W. Yao, H. Li, and H. Zhang, IEEE Access, **8**, 37124 (2020). https://doi.org/10.1109/ACCESS.2020.2975377
13. S. Gupta and R. R. Sedamkar, Int. J. Comput. Trends Technol. (IJCTT) **67**, 6 (2019). https://doi.org/10.14445/22312803/IJCTT-V67I6P109
14. J. P. Li, A. Ul Haq, S. Ud Din, J. Khan, A. Khan, and A. Saboor. IEEE Access, **8**, ID 107562 (2020). https://doi.org/10.1109/ACCESS.2020.3001149
15. E. Nasarian, M. Abdar, A. F. Mohammad, R. Alizadehsani, S. Hussain, E. B. Mohammad, M. Z. Moghadam, X. Zhou, I. P. PÅawiak, U. R. Acharya, R. S. Tan, and N. Sarrafzadegan, Pat. Recog. Lett. **133**, 33 (2020). https://doi.org/10.1016/j.patrec.2020.02.010
16. R. N. Abirami and P. M. D. R. Vincent, Cardiac Arrhythmia Detection Using Ensemble of Machine Learning Algorithms, in Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing, ed. K. Das et al. (Springer, Singapore, 2020) **1057**. https://doi.org/10.1007/978-981-15-0184-5_41
17. S. Gupta and R. R. Sedamkar, Genetic Algorithm for Feature Selection and Parameter Optimization to Enhance Learning on Framingham Heart Disease Dataset, in Intelligent Computing and Networking. Lecture Notes in Networks and Systems, ed. V.E. Balas et al. (Springer, Singapore, 2021) **146**. https://doi.org/10.1007/978-981-15-7421-4_2
18. B. A. Tama and S. Lee, BioMed. Res. Int. **2020**, ID 9816142 (2020). https://doi.org/10.1155/2020/9816142
19. M. Z. Uddin and M. A. Yousuf, J. Sci. Res. **7**, 11 (2015). https://doi.org/10.3329/jsr.v7i3.19527
20. S. Gupta and R. R. Sedamkar, Machine Learning with Health Care Perspective, Chapter: Machine Learning for Healthcare: Introduction, Learning and Analytics in Intelligent Systems (Springer Nature Book, 2020) **13**. https://doi.org/10.1007/978-3-030-40850-3
21. S. Yesmin, M. S. Huda, P. K. Biswas, M. I. I. Wahed, and T. Naz, J. Sci. Res. **8**, 81 (2016). https://doi.org/10.3329/jsr.v8i1.24375
22. M. R. Hasan and A. R. Baizid, J. Sci. Res. **9**, 67 (2016). https://doi.org/10.3329/jsr.v1i1.29308