Parameter Selection Algorithm for Building an Efficient Prediction Model using Machine Learning

Jasvant Dev S , Pavitharan R , Angelina Geetha, Abaid Isaac Ninan

Hindustan Institute of Technology and Science
Chennai, India

***Abstract*—In the current world, the range and size of data sets are enormous and to physically evaluate a data set is impossible. Here is where Machine Learning comes in handy. With the growing amount of ML algorithms, testing and training of data sets to predict a model has become straightforward. Furthermore, with the introduction of feature selection techniques, model prediction has become more efficient. In this system, we implement a customized wrapper method inside which we will find and remove the correlated features using a correlation algorithm on the given data set and then feed it into the wrapping algorithm that uses Sequential Forward Selection(SFS) technique thereby increasing the efficiency of the wrapping algorithm and the accuracy of the final predictive model. We implement three different prediction algorithms suchasDecision tree classification, Logistic regressionand Random forest classification on the selected feature sets of three different data sets. From our experimental results, it is proved that the performance of these standard algorithms differ considerably to different sets of data. Hence it is proved that the choice of suitable machine learning algorithm affects the accuracy and performance on a given data set.**

Keywords—Wrapper method, Correlation algorithm,Prediction algorithm, Machine learning

I. INTRODUCTION

In today's world, the range and size of datasets are enormous and to physically evaluate a data set is impossible. Here is where Machine Learning comes in handy. With the growing amount of ML algorithms, testing and training of datasets to predict a model has become straightforward.Predictive Analysis in ML has become the forefront in collecting current and historical data to forecast an activity.Since there are huge datasets with a high quantity of features, selecting the important features that are required for the model is highly important. Predictive selection solves this problem by selecting the required features of a model thereby saving time and resources. Although, implementing a feature selection algorithm on a highly populated data set takes immense amount of time to process and affects the accuracy of the predictive model.To solve this issue, In this system, we implement a customized wrapper method inside which we will find and remove the correlated features using a correlation algorithm on the given data set and then feed it into the wrapping algorithm for feature selection that uses Sequential Forward Selection(SFS) technique thereby increasing the efficiency of the wrapping algorithm and the accuracy of the predictive model. We implement three different prediction algorithms such asDecision tree classification, Logistic regression and Random forest classification on the selected feature sets of three different data sets to test and train the models and prove that each algorithm performs differently for different data sets.

II. LITERATURE REVIEW

Performance analysis of Machine Learning models was done by Raisa et al [1]. In their work, they indicated the implementation of GIWRF, which is known as Gini Impurity based Weighted Random Forest model as the enclosed feature selection and resulted with Decision Tree algorithm as the best performing after feature selection using Random Forest.

Salem [2] proposed the ensemble method which uses a bagging technique that improves parameter selection consistency in medical data sets using reduction in data variance. This method was successful in improving selection stability along with the ability to maintain the classification accuracy.

Christian Janiesch et al [3] proposed a classification model, that implements an algorithm which uses filter method for parameter selection using correlation and ML classifiers experiment. They have listed the four reasons for supporting the fact that feature selection is essential. Reducing the number of parameters, reduction in training time, minimal overfilling and minimizing the curse of dimensionality were the four points of concern.

Abid Ishaq et al [4] proposed a model that predicts patient's survival by employing nine classification algorithms and a technique called SMOTE for the imbalance class problem and uses Random Forest Classification for selecting the highly ranked features from the initial data set.

Rung-Ching et al [5] presented a comparative study of various predictive algorithms and chose Random Forest as a method to implement feature selection to classify multiple models and provides a relative study on the Random Forest algorithm from various perspectives.

Works of Shafiq et al [6] implemented a parameter selection algorithm named the CorrAUC that is evaluated using a wrapper technique which selects the best features for the model by implementing the AUC criterion. Then, it implements four different predictive algorithms for the prediction of the model.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

An Adaptive Feature Selection guided Deep Forest (AFS-DF) as an algorithm to detect COVID-19 based on the images of chest CT was proposed by Liang Sun et al [7]. It proposes a parameter selection technique that is derived from a trained deep forest model, thereby reducing the feature repetition.

Literature work of Joshua et al [8] proposes an optimal feature selection method to classify medical images by implementing a deep learning model which incorporates preprocessing of data set, parameter selection and classification. It proposes an Opposition-based Crow Search (OCS) algorithm that focuses on picking the optimal features to enhance the efficiency of the predictive model.

Yuyang et al [9] presented an intrusion detection system that is built upon the parameter selection and ensemble learning methods. It uses an algorithm called CFS-BA which helps in the reduction of dimensions thereby selecting the best subset of features by finding the feature correlation. Then it uses an ensemble method to classify and detect network intrusion.

The proposed work of Kazi et al [10] proposes a supervised ML system to classify and find whether a traffic in a network is legit or not. The system implements a supervised learning algorithm(ANN) along with wrapper based parameter selection technique to increase the performance of the predictive model.

An information based algorithm which logically chooses the best set of features for prediction was designed by Ambusaidi et al [11]. The proposed feature selection technique can execute both linear and non-linear data which are dependent. It uses this technique to select and classify the network intrusion detection.

Xuezhi Wen et al [12] proposed an effective feature selection technique which uses AdaBoost by merging a sample's parameter value to the label of its class. It uses an improved normalization algorithm on the chosen parameter values that is designed to improve the efficiency of feature selection and classification in vehicle detection.

Felipe and their team [13] propose an arrhythmia's detection algorithm which uses SVM classifiers. It uses a filter method based parameter selection technique to pick the optimal set of features required for the detection and classifies the model.

Principal Component Analysis (PCA) finds its application in various domains. Fengxi et al [14] implemented a feature selection method which uses Principal Component Analysis. It uses this method based on the perspective of numeral analysis and selects the optimal set of features for the face recognition predictive model.

Isabelle et al [15] presented a parameter selection technique which uses correlation analysis along with ML oriented methods to find optimal set of features from the given initial data set to improve the performance of the patient's mental health prediction model.

III. ALGORITHMS USED

A. CORRELATIONALGORITHM

Correlation is a technique that identifies the behaviour of one variable when compared to the other variable. It finds out the strength of correspondence between two or more variables. It is also known as a bi-variate data between two variables. Correlation is categorized into three types.They are,

1) Positive Correlation

a) If the value of one variable rises along with the value of other variable(s), it is called positive correlation.

2) Negative Correlation

b) If the rise in the value of one variablebrings about the decline in the value of other variable(s) or vice versa, it is called negative correlation.

3) No Correlation

c) If the rise or decline in the value of one variabledoes not cause any change in the value of other variable(s), it is called no correlation.

Benefits of implementing correlation algorithm:

- Based on the degree of correlation between two variables, one variable can be easily predicted from the second one.
- Implementing correlation algorithm helps in finding key variables and the variables that depend on the key variables.
- Over fitting can be reduced by the implementation of correlation algorithm.
- Data can be understood in a better way if analysis of correlation in done properly.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 20 August 2022

- Removing correlated features reduces the training time of the predictive model.
- By implementing correlation algorithm, we can remove the correlated features thereby increasing the accuracy of the predictive model.

In this system, we have implemented Pearson's correlation coefficient to find the correlated features.

B. WRAPPER METHODS

Wrapper methods are generally used for performing parameter selection using various algorithms. It uses a greedy search technique that calculates various combinations of parameters using the specified evaluation conditions. The evaluation condition can be specified as the measure of the performance that is based on the problem type.
There are three wrapper method techniques. They are,

1) *Forward selection*
2) *Backward elimination*
3) *Bi-directional elimination or Stepwise Selection*

C. FORWARD SELECTION

In this technique, we initially begin with a void set and start adding single feature each time to the set and select that feature which has the minimum probability value. Then we select the second feature by working out different combinations of features with the first feature and pick the feature that has the minimum probability value. Then we select the third feature by working out different combinations of features with the selected two features and pick the feature that has the minimum probability value. We repeat this method till we get a subset of features with the probability value of each feature being lesser when compared to the significance level.
Steps in forward selection are,

- Select a significance level of your own.
- Select the models by taking single feature at once. There can a total of y models. Pick that parameter which has the least probability value.
- Select the second feature by working out different combinations of features with the first feature.
- Then select that feature which has the least probability value. When probability value lesser than significance level, we move to the third step else we kill the execution.

D. BACKWARD ELIMINATION

In this technique, initially we begin with the entire feature set and start by removing individual features which has the maximum probability value. We continue this process till we gain a subset of parameters that are important.
Steps in backward elimination are,

- Select a significance level of your own.
- Take a model with the entire feature set.
- Start by removing individual features when the probability value is greater than the SL, else kill the execution.

E. BI-DIRECTIONAL ELIMINATION

In this technique, we add the individual features as in forward selection method but when appending a new parameter, it will compare the significance level of the previously added features with the new feature and removes those features that are insignificant with the help of backward elimination technique. Therefore, bi-directional elimination technique uses both forward selection technique and backward elimination technique.
Steps in bi-directional elimination are,

- Select a significance level of your own.
- Execute forward selection technique to select the individual features and add them to a subset.
- While adding individual features, if any of the previously added features has a significance level lesser than the newly added feature, execute backward elimination and remove that feature.
- Continue the above process till we arrive at an ideal subset.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

In our proposed system, we use theSequential Forward Selection(SFS**)** technique using Random Forest Classifier algorithm.

F. SEQUENTIAL FORWARDSELECTION

SequentialForward Selectionbegins with the estimation of individual features,thenpicks that feature that results in the top performing algorithm model. The best performing model entirely depends on the defined evaluation conditions. In the next step, all feasible combinations of the selected feature and a following feature are calculated and the second feature is selected. This process goes on continuously until the predefined number of features is established. In Sequential Forward Selection,features will be sequentially added to an empty set of features till the inclusion of extra features does not decrease the criterion.

G. RANDOM FOREST CLASSIFICATION

Random Forest Classification algorithm is a ML algorithm that is based on a technique called supervised learning. It is used in finding solutions forclassification as well asregression problems in machine learning. It uses the notion of ensemble learning. Ensemble learning is a process that can be implemented to enhance the execution of the model by bringing togethernumerous classifiers to find solution to complicated problems. As the name denotes, 'Random Forest' is a classifier algorithm which takes the subsets of the data set and puts them into multiple decision trees and further calculates and takes the average to improve the accuracy of the predictive model. Instead of completely relying on a single decision tree, the random forest algorithm calculates the average prediction from multiple decision trees and predicts the output.

H. DECISION TREE CLASSIFICATION

Decision Tree Classification algorithm uses a technique called supervised learning which is employedin solving classification as well asregression problems. But it is most commonly used in solving classification problems. It's a classifier that is structured like a tree. The internal nodes of the tree indicate the parameters of the data set, the branches of the tree indicates the branching conditions for the decision to be taken and the leaf nodes of the tree conveys the output of the predictive model. Decision Tree Classification follows a graphical representation to evaluate all the possible outcome of a problem using a set of specified conditions. The CART(Classification and Regression Tree) algorithm is used in order to assemble a tree. It functions in accordance with the Gini's impurity index. CART is a simple ML algorithm which can be used in various problems. In decision tree algorithm, based on the threshold value of an element,nodes are generally divided into sub-nodes. In relation to the Gini's impurity index, the Classification and Regression Tree algorithm employsit by lookingat the uniformityof the sub-nodes.The training set is nothing but the root node of the tree. Based on the finest element and the threshold value, the training set is split up into two parts. Additionally, with the similar logic, the subsets are divided. The above process is carried on until the final subset is established and the model is predicted.

I. LOGISTIC REGRESSION:

In Logistic regressionmethod, weconstruct predictive ML models when the target variable tends to be a binary variable. It's a classification model, which is easilyunderstandable and can attain greatexecution with directlydistinguishable classes. Generally, Logistic regression is used to identify the association between a target variable and other unrelated variables. Sigmoid function is where the logistic regression derives it's name from. The sigmoid function's rate stays between 0 and 1 and does not go past this boundary. Therefore, it creates a curve that looks like an 'S'. In this method, it employs a notion known as the threshold value. The threshold value determines the probability as a choice between 0 and 1. When the values end up over the threshold value, then it is considered as 1. When the values end up beneath the threshold value, then it is considered as 0.

IV. ARCHITECTURE DIAGRAM

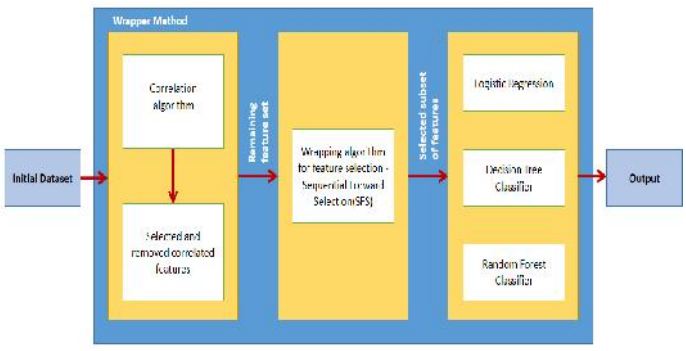

Fig. 1. Architecture diagram ofthe proposed system

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

V. DATASET

A. Company Bankruptcy Prediction Data set:

Data set source: kaggle
Data set name: company_bankruptcy.csv
Data set shape: 6819 rows x 96columns
Data set size: 31.41 MB
Data set format: Structured data
File format: CSV

B. PatientSurvival Prediction Data set:

Data set source: kaggle
Data set name: patient_dataset.csv
Data set shape: 91713 rows x 85 columns
Data set size: 11.46 MB
Data set format: Structured data
File format: CSV

C. Australian Weather PredictionData set:

Data set source: kaggle
Data set name: weatherAUS.csv
Data set shape: 142193 rows x 24 columns
Data set size: 14.17 MB
Data set format: Structured data
File format: CSV

VI. RESEARCH AND ANALYSIS

A. Company Bankruptcy Prediction Data set

Case: I

In this case, we take the entire feature set containing 96 columns and perform wrapping algorithm in it. A total of 27 features out of 96 features were correlated and we don't remove any of the correlated features. We then perform three different classification algorithms using 96 features to predict the accuracy of the model.
Number of correlated features: 27
Number of correlated featuresremoved : 0
Prediction Algorithm:

1) Logistic Regression

Confusion matrix: [[1301 15]
[480]]

TABLE I. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 1

	Precision	Recall	F1-score	Support
Not bankrupt	0.96	0.99	0.98	1316
Bankrupt	0.00	0.00	0.00	48

Accuracy: 95.38%

2) Decision Tree Classifier

Confusion matrix: [[1283 42]
[29 10]]

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

	Precision	Recall	F1-score	Support
Not bankrupt	0.98	0.97	0.97	1325
Bankrupt	0.19	0.26	0.22	39

TABLE II. PREDICTION RESULTS FOR DECISION TREE IN CASE 1

Accuracy:94.79%

3) Random Forest Classifier
Confusion matrix: [[13052]
[51 6]]

TABLE III. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 1

	Precision	Recall	F1-score	Support
Not bankrupt	0.96	1.00	0.98	1307
Bankrupt	0.75	0.11	0.18	57

Accuracy: 96.11%

Case: II

In this case, we take the entire feature set containing 96 columns and perform wrapping algorithm in it. A total of 27 features out of 96 features were correlated and we partially remove 15 correlated features from the 96 features. We then perform three different classification algorithms using the remaining 81 features to predict the accuracy of the model.

Number of correlated features: 27
Number of correlated features removed : 15
Prediction Algorithm:
1) *Logistic Regression*
Confusion matrix: [[1315 8]
[410]]

TABLE IV. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 2

	Precision	Recall	F1-score	Support
Not bankrupt	0.97	0.99	0.98	1323
Bankrupt	0.00	0.00	0.00	41

Accuracy: 96.40%

2) Decision Tree classifier
Confusion matrix: [[1285 39]
[26 14]]

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

TABLE V. PREDICTION RESULTS FOR DECISION TREE IN CASE 2

	Precision	Recall	F1-score	Support
Not bankrupt	0.98	0.97	0.98	1324
Bankrupt	0.26	0.35	0.30	40

Accuracy: 95.23%

3) Random Forest Classifier

Confusion matrix: [[13143]
[42 5]]

TABLE VI. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 2

	Precision	Recall	F1-score	Support
Not bankrupt	0.97	1.00	0.98	1317
Bankrupt	0.62	0.11	0.18	47

Accuracy: 96.70%

Case: III

In this case, we take the entire feature set containing 96 columns and perform wrapping algorithm in it. A total of 27 features out of 96 features were correlated and we completely remove all the the correlated features from the 96 features. We then perform three different classification algorithms using the remaining 69 features to predict the accuracy of the model.

Number of correlated features: 27
Number of correlated features removed : 27
Prediction Algorithm:

1) *Logistic Regression*

Confusion matrix: [[1319 4]
[41 0]]

TABLE VII. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 3

	Precision	Recall	F1-score	Support
Not bankrupt	0.97	1.00	0.98	1323
Bankrupt	0.00	0.00	0.00	41

Accuracy: 96.70%

2) Decision Tree Classifier

Confusion matrix: [[1288 35]
[24 17]]

TABLE VIII. PREDICTION RESULTS FOR DECISION TREE IN CASE 3

	Precision	Recall	F1-score	Support
Not bankrupt	0.98	0.97	0.98	1323
Bankrupt	0.33	0.41	0.37	41

Accuracy: 95.67%

3) Random Forest Classifier

Confusion matrix: [[13173]
[34 10]]

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

TABLE IX. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 3

	Precision	Recall	F1-score	Support
Not bankrupt	0.97	1.00	0.99	1320
Bankrupt	0.77	0.23	0.35	44

Accuracy: 97.28%

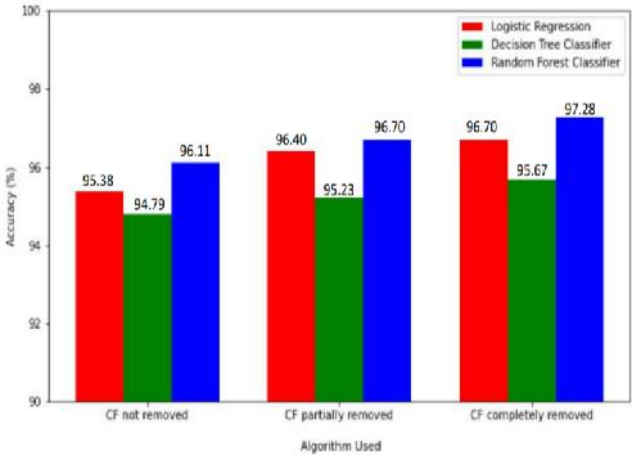

Fig. 2. Bar graph on results of company bankruptcy prediction

B. PatientSurvival Prediction Data set

Case: I

In this case, we take the entire feature set containing 85 columns and perform wrapping algorithm in it. A total of 24 features out of 85 features were correlated and we don't remove any of the correlated features. We then perform three different classification algorithms using 85 features to predict the accuracy of the model.

Number of correlated features: 24
Number of correlated features removed : 0
Prediction Algorithm:

1) Logistic Regression

Confusion matrix: [[10750 21]
[1049 30]]

TABLE X. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 1

	Precision	Recall	F1-score	Support
Not survive	0.91	1.00	0.95	10771
Survive	0.59	0.03	0.05	1079

Accuracy:90.97%

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

2) Decision Tree Classifier

Confusion matrix: [[9942 834]
[714 360]]

TABLE XI. PREDICTION RESULTS FOR DECISION TREE IN CASE 1

	Precision	Recall	F1-score	Support
Not survive	0.93	0.92	0.93	10776
Survive	0.30	0.34	0.32	1074

Accuracy:86.93%

3) Random Forest Classifier

Confusion matrix: [[1069587]
[827 241]]

TABLE XII. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 1

	Precision	Recall	F1-score	Support
Not survive	0.93	0.99	0.96	10782
Survive	0.73	0.23	0.35	1068

Accuracy: 92.28%

Case: II

In this case, we take the entire feature set containing 85 columns and perform wrapping algorithm in it. A total of 24 features out of 85 features were correlated and we partially remove 15 correlated features from the 85 features. We then perform three different classification algorithms using the remaining 70 features to predict the accuracy of the model.

Number of correlated features: 24
Number of correlated features removed : 15
Prediction Algorithm:

1) Logistic Regression

Confusion matrix: [[0773 27]
[1010 40]]

TABLE XIII. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 2

	Precision	Recall	F1-score	Support
Not survive	0.91	1.00	0.95	10800
Survive	0.60	0.04	0.07	1050

Accuracy: 91.24%

2) Decision Tree Classifier

Confusion matrix: [[9975 791]
[692 392]]

TABLE XIV. PREDICTION RESULTS FOR DECISION TREE IN CASE 2

	Precision	Recall	F1-score	Support
Not survive	0.94	0.93	0.93	10766
Survive	0.33	0.36	0.35	1084

Accuracy: 87.48%

3) Random Forest Classifier

Confusion matrix: [[10717117]

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

[768 248]]

TABLE XV. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 2

	Precision	Recall	F1-score	Support
Not survive	0.93	0.99	0.96	10834
Survive	0.68	0.24	0.36	1016

Accuracy: 92.53%

Case: III

In this case, we take the entire feature set containing 85 columns and perform wrapping algorithm in it. A total of 24 features out of 85 features were correlated and we completely remove all the the correlated features from the 85 features. We then perform three different classification algorithms using the remaining 61 features to predict the accuracy of the model.

Number of correlated features: 24
Number of correlated features removed : 24
Prediction Algorithm:

1) Logistic Regression

Confusion matrix: [[10824 23]
[977 26]]

TABLE XVI. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 3

	Precision	Recall	F1-score	Support
Not survive	0.92	1.00	0.96	10847
Survive	0.53	0.03	0.05	1003

Accuracy:91.56%

2) Decision Tree Classifier

Confusion matrix: [[10041 784]
[669 356]]

TABLE XVII. PREDICTION RESULTS FOR DECISION TREE IN CASE 3

	Precision	Recall	F1-score	Support
Not survive	0.94	0.93	0.93	10825
Survive	0.31	0.35	0.33	1025

Accuracy: 87.73%

3) Random Forest Classifier

Confusion matrix: [[1076180]
[784 225]]

TABLE XVIII. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 3

	Precision	Recall	F1-score	Support
Not survive	0.93	0.99	0.96	10841
Survive	0.74	0.22	0.34	1009

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

Accuracy: 92.70%

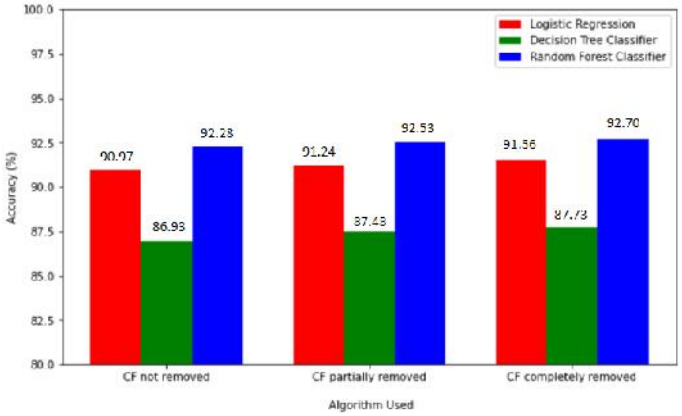

Fig. 3. Bar graph on results of patient survival prediction

C. Australian Weather PredictionData set

Case: I

In this case, we take the entire feature set containing 24 columns and perform wrapping algorithm in it. A total of 3 features out of 24 features were correlated and we don't remove any of the correlated features. We then perform three different classification algorithms using 24 features to predict the accuracy of the model.

Number of correlated features: 3
Number of correlated features removed : 0
Prediction Algorithm:

1) Logistic Regression

Confusion matrix: [[8659 108]
[359 2158]]

TABLE XIX. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 1

	Precision	Recall	F1-score	Support
Not rain	0.96	0.98	0.97	8767
Rain	0.95	0.86	0.90	2517

Accuracy: 95.86%

2) Decision Tree Classifier

Confusion matrix: [[8837 0]
[0 2447]]

TABLE XX. PREDICTION RESULTS FOR DECISION TREE IN CASE 1

	Precision	Recall	F1-score	Support
Not rain	1.00	1.00	1.00	8837
Rain	1.00	1.00	1.00	2447

Accuracy: 100%

3) Random Forest Classifier

Confusion matrix: [[87670]
[0 2517]]

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

TABLE XXI. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 1

	Precision	Recall	F1-score	Support
Not rain	1.00	1.00	1.00	8767
Rain	1.00	1.00	1.00	2517

Accuracy: 100%

Case: II

In this case, we take the entire feature set containing 24 columns and perform wrapping algorithm in it. A total of 3 features out of 24 features were correlated and we partially remove 1 correlated feature from the 24 features. We then perform three different classification algorithms using the remaining 23 features to predict the accuracy of the model.

Number of correlated features: 3
Number of correlated features removed : 1
Prediction Algorithm:

1) Logistic Regression
Confusion matrix: [[8664 123]
[239 2258]]

TABLE XXII. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 2

	Precision	Recall	F1-score	Support
Not rain	0.97	0.99	0.98	8787
Rain	0.95	0.90	0.93	2497

Accuracy: 96.79%
2) Decision Tree Classifier
Confusion matrix: [[8794 0]
[0 2490]]

TABLE XXIII. PREDICTION RESULTS FOR DECISION TREE IN CASE 2

	Precision	Recall	F1-score	Support
Not rain	1.00	1.00	1.00	8794
Rain	1.00	1.00	1.00	2490

Accuracy: 100%

3) Random Forest Classifier
Confusion matrix: [[8803 0]
[0 2481]]

TABLE XXIV. PREDICTION RESULTS FOR RANDOM FOREST IN CASE 2

	Precision	Recall	F1-score	Support
Not rain	1.00	1.00	1.00	8803
Rain	1.00	1.00	1.00	2481

Accuracy: 100%

Case: III

In this case, we take the entire feature set containing 24 columns and perform wrapping algorithm in it. A total of 3 features out of 24 features were correlated and we completely remove all the the correlated features from the 24 features. We then perform three different classification algorithms using the remaining 21 features to predict the accuracy of the model.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

Number of correlated features: 3
Number of correlated features removed : 3
Prediction Algorithm:

1) Logistic Regression
Confusion matrix: [[8752 103]
[168 2261]]

TABLE XXV. PREDICTION RESULTS FOR LOGISTIC REGRESSION IN CASE 3

	Precision	Recall	F1-score	Support
Not rain	0.98	0.99	0.98	8855
Rain	0.96	0.93	0.94	2429

Accuracy: 97.59%

2) Decision Tree Classifier
Confusion matrix: [[8765 0]
[0 2519]]

TABLE XXVI. PREDICTION RESULTS FOR DECISION TREE IN CASE 3

	Precision	Recall	F1-score	Support
Not rain	1.00	1.00	1.00	8765
Rain	1.00	1.00	1.00	2519

Accuracy: 100%

3) Random Forest Classifier
Confusion matrix: [[8755 0]
[0 2529]]

TABLE XXVII. PREDICTION RESULT FOR RANDOM FOREST IN CASE 3

	Precision	Recall	F1-score	Support
Not rain	1.00	1.00	1.00	8755
Rain	1.00	1.00	1.00	2529

Accuracy: 100%

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 20 August 2022

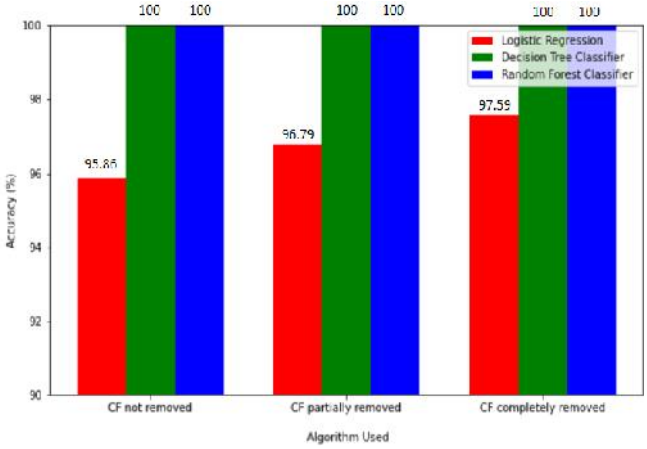

Fig. 4. Bar graph on results of Australian weather prediction

VII. CONCLUSION AND FUTURE

From the above evaluated results, we found out that for broader datasets, implementation of feature selection plays a major role in the efficiency of the predictive model. By removing the correlated features before implementing the predictive algorithms, we can improve the efficiency of the wrapping algorithm thereby improving the accuracy of the predictive model. From the above evaluated results, for the Patient Survival Prediction data set, RandomForest Classifier displayed the best results. For the Company Bankruptcy Prediction data set, RandomForest Classifier again displayed the best results. For the Australian Weather Prediction dataset, both Decision Tree Classifier and RandomForest Classifier displayed the best results. In future, we will look at implementing the same method to much broader data sets to yield much better performance of the predictive models.

REFERENCES

[1] RaisaAbedin Disha and SajjadWaheed, "Performance analysis of machine learning models for intrusion detection system using Gini impurity-based weighted Random Forest (GIWRF) feature selection technique," Cybersecurity 5:1, 2022.

[2] Salem Alelyani, "Stable bagging feature selection on medical data," 8:11,Journal of Big Data, 2021.

[3] Christian Janiesch, Patrick Zschech and Kai Heinrich, "Machine learning and deep learning," Electronic Markets (2021) 31:685–695, April 2021.

[4] Abid Ishaq, Saima Sadiq, Muhammad Umer, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara and Michele Nappi, "Improving the prediction of heart failure patient's survival using SMOTE and effective data mining techniques," IEEE access 9. 39707-39716, 2021.

[5] Rung Ching Chen, Christine Dewi, Su Wen Huang and Rezzy Eko Caraka, "Selecting critical features for data classification based on machine learning methods," 7:52, Journal of Big Data, 2020.

[6] Muhammad Shafiq, Zhihong Tian, Ali Kashif Bashir, Xiaojiang Du and Mohsen Guizani, "CorrAUC:a malicious bot-IoT trafic detection method in IoT network using machine-learning techiques," IEEE internet of things journal 8(5),3242-3254, 2020.

[7] Liang Sun, Zhanhao Mo, Fuhua Yan, Liming Xia, Fei Shan, Zhongxiang Ding, Bin Song, Wanchun Gao, Wei Shao, Feng Shi, Huan Yuan, Huiting Jiang, Dijia Wu, Ying Wei, Yaozong Gao, He Sui, Daoqiang Zhang and Dinggang Shen, "Adaptive feature selection guided deep forest for covid-19 classification with chest ct," IEEE journal of Biomedical and Health informatics 24(10),2798-2805, 2020.

[8] Joshua Samuel Raj, S Jeya Shobana, Irina Valeryevna Pustokhina, Denis Alexandrovich Pustokhin, Deepak Gupta and K Shankar, "Optimal feature selection-based medical image classification using deep learning model in internet of medical things," IEEE Access 8,58006-59017, 2020.

[9] Yuyang Zhou, Guang Cheng, Shanqing Jiang and Mian Dai, "Building an efficient intrusion detection system based on feature selection and ensemble classifier," Computer networks 174, 107247, 2020.

[10] Kazi Abu Taher,Billal Mohammed Yasin Jisan and Md Mahbubur Rahman, "Network intrusion detection using supervised machine learning technique with feature selection," International conference on robotics,electrical and signal processing techniques(ICREST),643-646,2019.

[11] Mohammed A Ambusaidi,Xiangjian He, Priyadarsi Nanda and Zhiyuan Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," IEEE transactions on computers 65(10),2986-2998,2016.

[12] Xuezhi Wen, Ling Shao, Wei Fang and Yu Xue, "Efficient feature selection and classification for vehicle detection," IEEE Transactions on Circuits and system for video technology 25(3),508-517,2014.

[13] Felipe Alonso-Atienza, Eduardo Morgado, Lorena Fernandez-Martinez, Arcadi Garcia-Alberola and José Luis Rojo-Alvarez, "Detection of life-threatening arrhythmias using feature selection and support vector machines," IEEE Transactions on biomedical Engineering 61(3),832-840,2013.

[14] Fengxi Song,ZhongweiGuo andDayong Mei, "Feature selection using principal component analysis," International conference on system science, engineering design and manufacturing information 1,27-30,2010.

[15] IsabelleGuyon and Andre Elisseeff , "An introduction to variable and feature selection," Journal of Machine Learning Research 31157-1182, March 2003.