How to Cite:
Amutha, P., & Priya, R. (2022). An optimized random forest for multi class classification to classify the students using machine learning approaches. *International Journal of Health Sciences*, *6*(S7), 5312-5326. https://doi.org/10.53730/ijhs.v6nS7.13116

An optimized random forest for multi class classification to classify the students using machine learning approaches

P. Amutha
Research Scholar, Dept. of ComputerScience VISTAS, India
Email: amuthasenthilp@gmail.com

Dr. R. Priya
Professor, Dept. of ComputerScience VISTAS, India
Email: priyaa.research@gmail.com

***Abstract*---**In recent years, Multi-class classification in Educational Data Mining (EDM) has been continued to be a focusing research to deal the issues in the imbalanced data set and less number of unifying classification algorithms. Machine-learning algorithms take prominent part in multi-class classification. This research studyintroduced an Optimized Random Forest for Multi-Class Classification (ORFMCC) to classify the students based on the higher education programs enrolment. The base classifier Random Forest is optimized by hyper-parameter tuning and feature selection processes. The Optimized RFMCC is developed in Python 3.3 using Spider IDE 4.1.5. The optimum parameter extracted by tuning and relevant features significantly improved classification accuracy. The classification performance of Optimized Random Forest for Multi-Class Classifieris compared with RF, NB, DT, LR, and KNN. The experiments result of Accuracy, F1-Score, precision and Recall revealed that the ORFMCC outperformed in Multi-Class classification compared to the other five Classifiers.

***Keywords*---**Multi Class, Hyper-Parameter, Feature Selection, Feature Important Score, Random Forest.

Introduction

EDM (Educational Data mining) is an emerging research area that involves statistical methods, methods in data mining as well as machine learning to analyze the different data in educational institutions and impart the hidden data for further decision-making with better accuracy. Data mining- Machine learning in the EDM field opens up new ideas that give new challenges for scientists and

researchers and enable the creation of positive interactions with different parts of the education system. Several popular techniques such as sequential pattern, clustering, prediction, classification, and association rule analysis are available to identify the problems in the education field [18]. Classification is one ofthe most focused data mining techniques adopted by the researcher to classify and predict data effectively and efficiently.

Classification is a supervised learning method used to classify the class labels in the data set. Generally, four types of classification techniques are used in data mining–machine learning including Binary (Classify the data belonging absolutely to one of the class two class labels), Multi-Class (Classify three or more classes), Multi-Label (Classify the data instances that can fit none of the classes or all these classes), and Imbalanced Classifications (where the data sets contain uneven of classes).

There are limited educational data mining (EDM) researches focusing on multi-class classification. The key process of this research is to introduce an 'Optimized random Forest for Multi-Class classification to classify the higher education courses for the higher secondary student's enrolment in higher education institution'.

I. Related Work

Dealing with Misclassification and over-fitting may be a serious problem in the machine learning era. Bujang, SitiDianah Abdul, et al. [1] proposed a model for multi-class prediction to predict final grades for the students from their earlier semester performance. This research reduced the overfitting and misclassification by introducingthe combination of two feature selection techniquesand the 'Synthetic Minority Oversampling Technique (SMOTE)'. SMOTE with modified parameter nearest neighbor was considered to deal with the imbalanced data set and attributes selection carried out by attribute evaluator wrapper (WFSJ4.8) with best-fit search (BFS) and FSJ4.8 with Ranker search. The result revealed that random forest with the proposed predictive model had significant improvement in the final prediction, and this proposed model showed enhanced prediction performance in imbalanced multi-class classification for students' grade prediction with five class labels.

Tripathi, Ankita, et al proposed multi-class random (MCRF) forest to classify small peptides of categories. The MCRF introduced using a ensemble random forest to handle the unbalanced data and approach for multi-class namely one versus all to deal with multi-class data. The developed model was trained with m classes and m random forest to achieve reliability by experimenting with various split applied for training-testing and 10 cross-validations applied to evaluate the result shown by the model. The statistical test kappa statistics and Wilcoxon sign rank were also applied to prove the goodness of the model which has a statistics value of more than 0.6. The research proved that the MCRF classifier was well suited for multiclass supervised data [2].

A survey was conducted to analyze the awareness of higher education courses amongst higher secondary students. The survey focused on the student's future

opportunities depending on the successful completion of a student's higher education degree course. The students who have insufficient knowledge to choose a suitable degree course need appropriate counselling about their higher education. Further, students from poor socioeconomic and rural areas pursued non-academic programs due to their lack of knowledge. The survey revealed that the efficient recommendation model has played as a counselor for the students to select comfortable courses based on students' skills and intentions [3].

Sevastyanov, Leonid A., and Eugene Yu Shchetinin [4] developed an effective algorithm to deal with over-fitting in case of imbalanced conditions in the data set. The crucial aim of this research isto increase accuracy in multi-class problems by eliminating the class imbalance. This proposed algorithm used feature selection methods namely recursive exclusive RFE, decision tree RF with feature importance, and Boruta to select significant attributes for classification. Additionally, sampling methods random sampling, SMOTE, and ADASYN were applied for converting class imbalance into a balance class. The result concluded that the combination of random forest feature selection with the ADASYN sampling method outperformed the other methods.

Hassan H et al**.** developed a hybrid method to overcome the problem related to the imbalanced multi-class classification. There are three sampling techniques applied and tested on five ensemble classifiers. The result revealed in the experiment, that the hybrid method ROS with AdaBoost produced excellent performance than other ensemble classifiers. Among seven sampling techniques, SMOTEENN (SMOTE Edited Nearest Neighbors) with ensemble classifiers constantly produced high results in students' performance prediction [5].

The conceptual framework was created for higher-secondary students' enrolment in higher education degree courses. This framework provided a new perspective to students' community to select their academic path and avoid dropping out of their enrolled degree. They concluded that the students need in-depth knowledge about higher education courses for a smooth transition from higher secondary to higher education and successful completion of their degree courses[8].

Dalton Ndirangu et al. developed a heterogeneous ensemble model to classify multiclass data sets and detect outliers by combining ensemble classifiers. In the preprocessing stage, outliers detected by global outliers and datasets are resampled using the synthetic minority oversampling technique. Binarization is applied on datasets using OneVsOne and the ensemble model is built using the base classifiers AdaBoost, random subspace algorithms and random forest. The 10-fold stratified cross-validation had enforced to evaluate the proposed model's performance. The research concluded that proper preprocessing and decomposing of multiclass data sets could improve the accuracy in minority outlier classes, and the integrity of majority classes was also maintained [10].

For classifying Arrhythmia Patients, the new model was created using a Super vector Machine based on multiclass classification. The improved feature selection techniques using wrapper method built on random forest classifier employed to select distinguished attributes and reduce the dimensionality of the data sets. SVM approaches 'one-against-one (OAO), one-against-all (OAA), an error-

correction code (ECC)' were applied to the selected feature set and categorized the patients into 'sixteen subclasses of arrhythmia'. The research revealed that OAO with SVM produced an impressive performance in terms of 'accuracy, kappa statistics, and root mean square error' [11].

An algorithm implemented to improve the accuracy of the weak algorithms using ensemble classification on a medical data set. In this research, the outputs of the multiple classifiers were combined to calculate the accuracy of the weak classifiers. This proposed ensemble algorithm employed bagging and boosting methods to predict heart disease at an initial stage and also achieved some improvement in accuracy. The experiment revealed that accuracy improved at most 7% for weak classifiers. Further, accuracy is enhanced using feature selection along with ensemble classification methods [13].

Two generalized multiclass uMRBBag (under-sampling RBBag) and oMRBBag (oversampling RBBag) were proposed variation of original bagging and a single J4.8 decision tree as multi-class baseline classifiers. In nature, the Roughly Balanced Bagging base classifier is unable to handle multi-class. So RBBag was modified slightly (modified RBBag named as RBBag*) to permit act as a base learner for multi-class classification and also applied modified bootstrap sampling method in modified RBBag* to avoid dataset for learning into a set of minority and majority samples. The stratified sampling was applied to consider the summation of all the minority classes instead of considering the size of minority classes to enable RBBag to treat multi classes. The geometric mean was taken for the performance evaluation [14].

AshishDutt and MaizatulAkmar Ismail introduced the method to improve the classifier's performance and reduce a fewer rate by using dimensionality reduction. In this research, the dimension of the data set was reduced by applying feature selection and multi-class data manipulated using the Learning Vector Quantization algorithm to recognize significant predictors for reducing the bias. The performance of the dimensionality reduction compared with Discriminate Analysis (LDA), Classification and Regression Tree (CART), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF). The result had shown that Random Forest with LVQ produced optimum results [15].

The KNN-based classification algorithm was introduced for sentiment identification and classifying the Twitter data into seven classes. In this proposed algorithm, distance-weighted KNN is employed to assist the weight of each attribute that is monitored. The attributes' weight can be a greater impact on determining the suitable samples' classes. The range between 0-1 is typically assigned for weights and value zero was assigned for irrelevant attributes. The process of assigning weight to the attributes plays a great role in the classification process. Also, it reduces the cost of the computation by introducing the indexing in training data sets like KD Tree and Ball Tree for large data sets [16].

Katuwal, Rakesh, and Ponnuthurai N. Suganthan introduced a variation of the oblique decision tree-based linear classifier using the ensemble method to enhance the performance on several multiclass data sets. Ensemble classifier wasbuilt around merging neural network, random vector functional link network

and oblique decision trees. The usage of random vector functional link networks had taken reduced training time. Using a neural network each training bag is partitioned into the number of subsets of classes and decision trees based on subset partition were used to train each partition. This proposed method demonstrated high performance compared with other classifiers [17].

Wang, Ying, et al [19] introduced a text-based ensemble classifier to identify the incidents by their type and severity. They used OVO and OVA ensemble strategies to transform the multi-class problem into several binary class problems. For experiments regularized logistic regression, linear support vector machine, and SVM with a radial basis function (RBF) kernel were considered and tested on benchmark data sets. One-versus-one ensemble of binary SVM RBF classifiers with binary count feature extraction was the effective method amongst others.

Maryam, et al, developed a model for diagnosis to predict erythemato-squamous disease using "multi-class SVM and hybrid method of feature selection" [20]. For feature selection, Chi-Square and Genetic algorithm are used along with filter and wrapper methods to select relevant features from the original features in the data set. In Feature selection, Chi-square is considered as a filter method to remove the redundant attributes and the Genetic algorithm is considered as a wrapper feature selection method to extract the ideal feature subset with SVM classifier. The 10-fold cross-validation was applied to evaluate the experimental result and the result revealed that the proposed SVM using multiclass combined with Chi-Square and GA produced better accuracy with 18 optimum features.

Chaudhary, Archana, et al, developed an improved random forest for multi-class classification using an attribute evaluator and an instance filter resample method along with Random forest to improve the accuracy of the multiclass groundnut disease dataset. For attribute selection, Correlation-based feature selection, Symmetrical uncertainty, and Gain ratio are used to select relevant attributes. Additionally, a Simple random sampling instance filter method was applied for balancing the class distributions of the multiclass dataset. The performance of the improved random forest was compared with the random forest with respect to the accuracy, F-measure, sensitivity, ROC, and specificity [23]. The performance of this improved random forest was tested on five benchmark datasets and the result revealed that the improved random forest outperformed in all these data sets.

Two Independent algorithms using hybrid mining were introduced to increase the accuracy rate for multiclass classification using decision tree and naïve Bayes. In a hybrid decision tree algorithm, naïve Bayes was employed to eliminate the noise instances from the dataset that assigned as training set because DT suffers overfitting and decrease accuracy due to irrelevant noisy instances in the data set. Next, a hybrid NB classifier was developed by using Decision Tree Induction to extract the most significant subset of attributes and the attribute weighting method which acts as an essential role in the final Classification. Performances of both methods were evaluated on ten benchmark data sets and 10- fold cross-validation was applied to estimate the output. The result showed that both proposed methods were best performed in real-time multiclass classification [24].

III. Methodology

Higher education program enrolment is the toughest task for higher secondary school students.The key process of this proposed work is to create a prediction algorithm to find the higher education program and to develop an Optimized Random Forest for Multiclass Classification (Henceforth ORFMCC) to handle multi-class using classification technique in Data mining- Machine learning to classify the students based on the most significant features that extracted during the features selection. The framework of the proposed methodology is depicted in Figure 3.1.

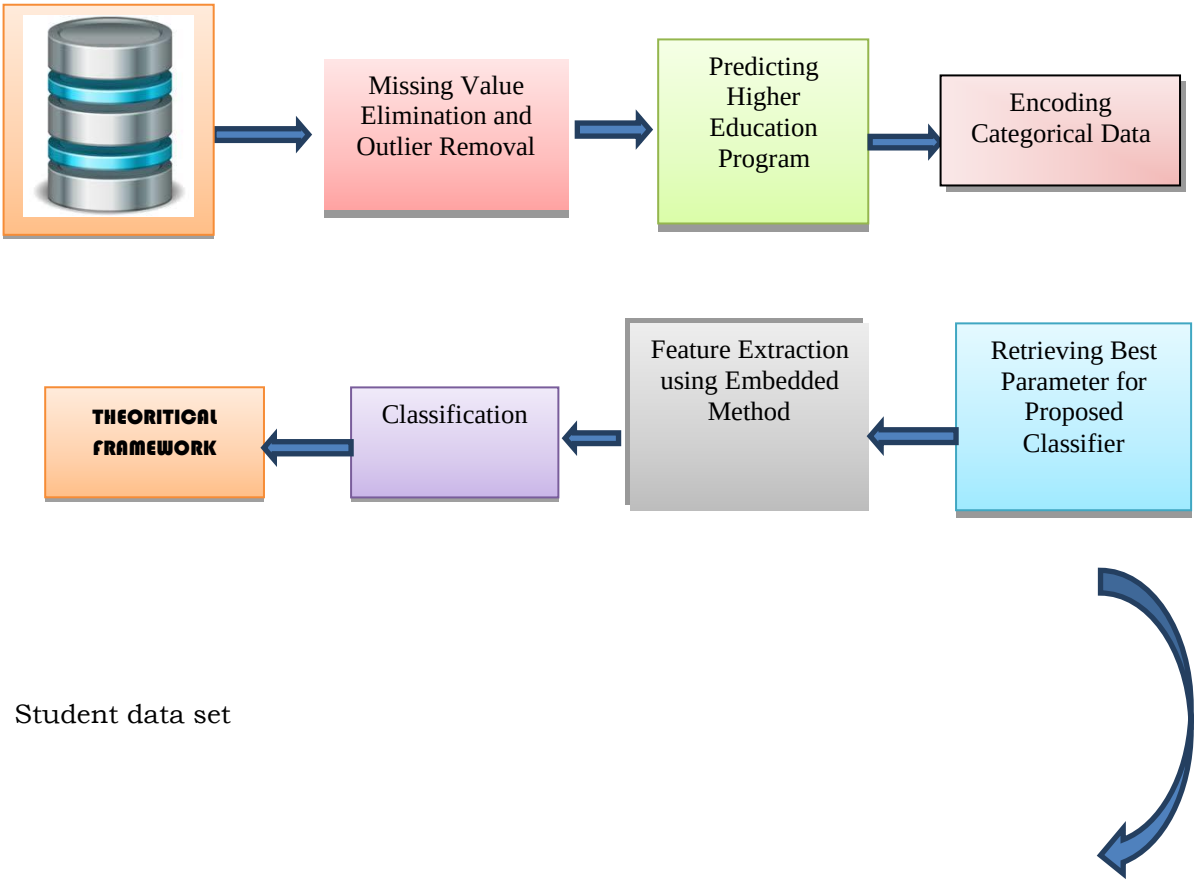

Figure 3.1 Framework for the proposed methodology to predict higher education programs and classify the student

The proposed work comprises the following three phases.

Phase I: It focuses on data preprocessing, finding higher education programs and ConvertingCategorical data

Phase II: It Focuses on classifying students based on the most significant features in the Student dataset
Phase III: It focuses on comparison among the classifiers using classification metrics.

Phase I: Data Pre-Processing, Finding Higher Education Program and Data Encoding

Phase I focuses on data preprocessing and finding the target variable. Student data is used for this research work. Data were collected using a well-defined Questionnaire. The questionnaire isprepared from previous research work and suggestions have been given by the academicians and incorporated. Initially, the data set consists of 33 attributes with no target variable and 925 instances. Most of them are categorical data and each has more than two options. The description of the student dataset is shown in Table 3.1.

Table 3.1 Description of Student's data set

S.NO	Description	Type	No of Values
1	20 Attributes	Object	More than two options
2	8 Attributes	Boolean	Two options
3	Target variable ABBHEC	Object	54 class labels

Handling Missing Values and Outliers

Data in the real world have missing values. Sometimes missing values are provided in one or more features. The missing value is a big problem in data mining – machine learning algorithms. If not rectified, some of the classifiers in machine learning create bias in results meanwhile decision trees and random forests can handle missing values. There are two ways to rectify the problem with missing values namely list-wise deletion and imputation. In listwise deletion, row deletion can be performed when few missing values in a row otherwise column deletion can be performed when too many missing values in an attribute [12]. In imputation, the replacement of median or mean value can be substituted for numerical data and the mode value can be replaced for nominal data.

Student data have missing values in features; these missing values were detected and rectified using the imputation method. In this research work, missing values are detected in the attributes namely category, m_income, family_enroll, location, extra_curri, medium_study, efficiency, entrance_cleared, problemsol_skill, and listening skill. These missing values were replaced by a most frequent value of an attribute in the dataset.

Outlier is another problem that affects the performance of the classifiers. Most of the previous research only focused on detecting outliers in numerical data. They applied the various outliers' methods which are handling numeric, discrete and

continuous, etc., [6][7][22]. For handling outliers in categorical data, the option value in an attribute was monitored and also considered as an outlier if the frequency distribution of an option in a particular attribute is less than 5 percent. These outliers are also reduced by replacing the most frequent value. The attributes in the student dataset namely purpose, finance_study, guidedby, efficiency, category, and m_income are having fewer values in their options. These values are treated as outliers and also substituted by the mode of the corresponding option in its attributes. The output of the proposed classifier is significantly improved by removing outliers.

Course Prediction

Features in the data set consist of students' personal information, demographic, school detail, intention to enroll in higher education, and their skills such as Communication, Thinking, Reasoning, Organization/Planning, Listening, Leadership, Writing, Programming, Decision making, etc. For supervised Machine Learning classification, the target variable plays a vital role in classification. The course prediction part in Phase I focuses on creating a course prediction algorithm to find the target variable. The algorithm for finding higher education programs (target variable) is as follows.

Algorithm 1: Algorithm for findinghigher education program

Input: Student Data set and Skill Data Set

Output: The predicted higher Education courses ABBHEC.

1. Begin
2. Read data sets
3. Compare student's Higher Secondary Group, Intention and Skills in Students data set with the Skill data set
4. Find Higher Education Courses based on Matching found in Step3
5. If more than one course is found for each student then
    a. Extract the first letter from the Final course.
    b. Join extracted letter and form a single abbreviated term
    c. Assign abbreviated terms into the Variable ABBHEC for each student.
6. End

For finding the target variable there are two data sets used namely the student data set and skill data prescribed by various academicians. The skill data set consists of a group of studies (PCB, PCM, and PCMB) and skills recommended by the academicians for the higher education course enrolment, and its

corresponding degree courses. In this research, 40 courses are considered initially. The description of the Skill dataset is shown in Table 3.2

Table 3.2 description of skill data set

S.No	Description	Type	No of Values
1	1 Attribute	Object	4 options
2	1 Attribute	Object	3 options
3	8 Attributes	Boolean	2 options
4	HEC	Object	40 options

There are various factors can affect the performance of a machine learning model. One factor that determines the performance of a model is how the data is processed and fed to the model. Therefore, coding the data is an important process because it transforms the data into categorical variables that the machine learning model can understand. Encodingthe data improves the quality of the model and is useful for feature engineering [9][21].

This study, the categorical features in the data set with more than two options were considered for data encoding process. The most popular categorical encoding methods are one-hot encoding and Label encoding which are considered for converting categorical attributes in the student data set. One-hot encoding creates high dimensionality compared to label encoding. When applying Label encoding, it does not create high dimensionality of data instead only assigns a numerical value from the 0-maximum option in a particular column. These converted data were sent through various classifiers RF, KNN, LR and NB, outperformed encoding technique has been chosen for further optimization to classify the student. For selecting outperformed classifier on encoded data, comparisonsamong classifiers were performed with respect to the accuracy, F1-score, recall and precision with less training time consumption.

The experiment determined that the Random Forest with Label encoding outperforms the other classifier for classifying multiple class labels in the students' data and also it consumes less time for training the data during the learning process.

Phase II: Classification

In this phase, the proposed classifier namely Optimized Random Forest for Multi Class Classification (ORFMCC) developed to attain high accuracy compared to the existing classifiers introduced by the various authorsRandom forest is an efficient method to classify multi classes in the output variable (Target variable). In this phase, the random forest has been considered as a base classifier and increases the accuracy of the base classifier; embedded feature selection was performed to

extract the most significant features using feature important score during the training of the classifier.
The feature selection process in the proposed framework extracted the most significant features which have more impact on classification. The relevant features are the stream of study, student's intention, various skills comm_skill, thinking_skill, problemsol_skill, org_plan_skill, listening_skill, leadership_skill, writing_skill, programming_skill and decision_making_skill. These features are selected using the feature's importance score that meet the given threshold in the learning process. Hyperparametertuning was performed using RandomizedSearchCV and 3, 5, and 10-cross validation is applied to select the best parameter for improving accuracy. The parameters considered for hyper tuning are max_features, n_estimators, min_samples_splitand max_depth. Finally, classifications are performed by extracting features along with optimum parameters on 70 percent of training data and 30 percent of test data.

Phase III: Evaluation

In this phase, the Performance of the Optimized Random Forest for multi-class classification (ORFMCC) is compared with five classifiers namely Random forest, Decision tree, Super vector machine, Logistic regression, K-nearest neighbors, and Naïve Bayes. The metrics Accuracy, f1-score, recall, and precision are calculated for performance comparison.

IV. Results and Discussion

In the proposed research work, preprocessing, feature selection, hyperparameter tuning and machine learning approaches are carried out using Python 3.3 SciKit Learn packages under Spider IDE. The system specification is Dell Inspiron, 12 GB RAM, Intel(R) Core i5- 5200U and 64 bit Windows 10.

The proposed Optimized Random Forest for multi-class classification is compared with five advanced classifiers with reference to accuracy, f1-score, precision and recall. Table 4.1 shows the performance metrics of classifiers namely 'Random Forest (RF), Decision Tree (DT), K- Nearest- Neighbors (KNN), Logistic Regression (LR), Naïve Bayes (NB) and proposed ORFMCC'.

Table 4.1 Output of the Classification metrics

CLASSIFIERS	ACCURACY	RECALL	PRECISION	F1-SCORE
RF	90.66	90.46	89.48	89.32
DT	94.60	94.6	95.86	94.49
KNN	43.00	43.00	49.90	43.50
LR	73.20	73.20	79.70	72.40
NB	83.70	83.70	90.10	83.20

ORFMCC(Proposed)	**97.48**	**97.83**	**97.96**	**97.31**

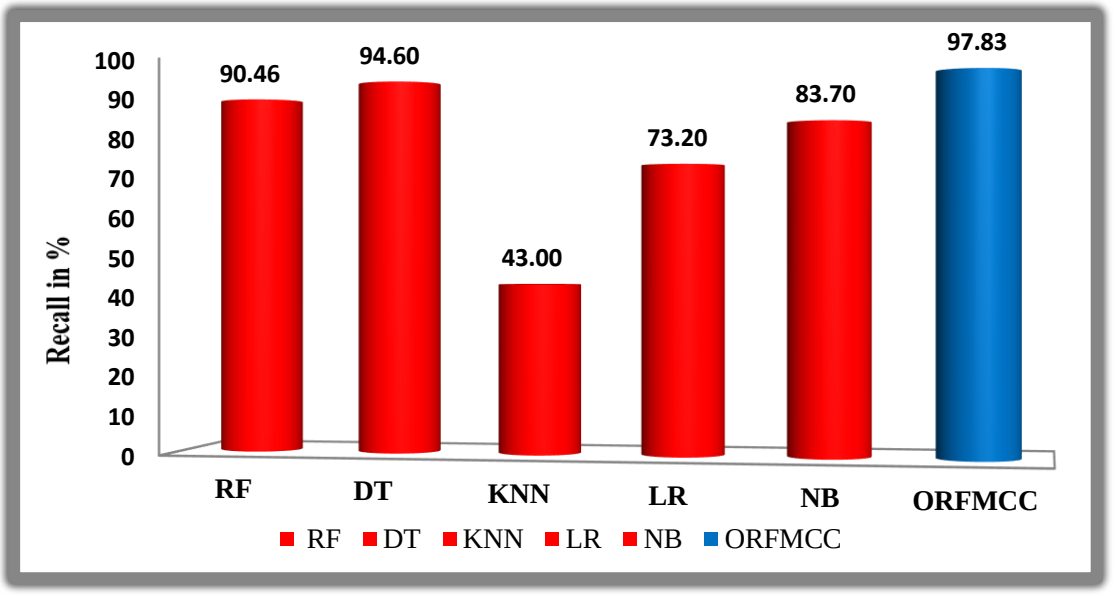

Figure 4.5 shows the classification accuracy of the Optimized Random Forest and the other five classifiers. The accuracy represents the percentage of correct predictions during classification. The result shows that the Optimized random Forest performed better than other classifiers

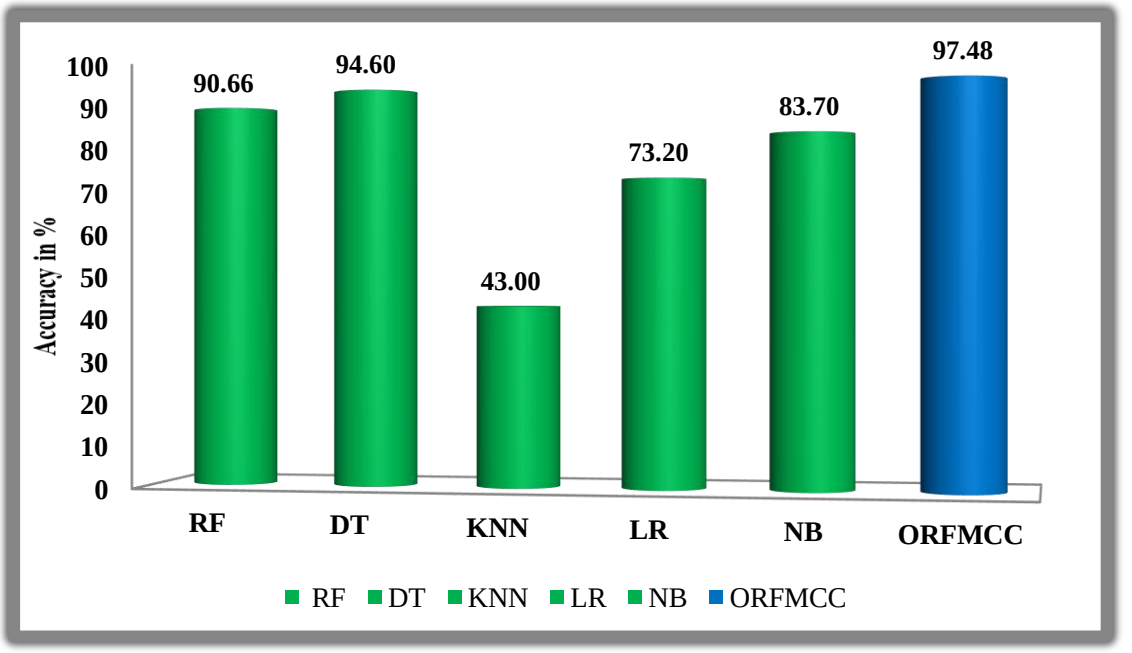

Figure 4.5 Classification Accuracy
Figure 4.6 Recall

Figure 4.6 shows the recall which represents the fraction of correctly predicted positive observations among other observations in the class. Here, the proposed ORFMCC shows better results than others.

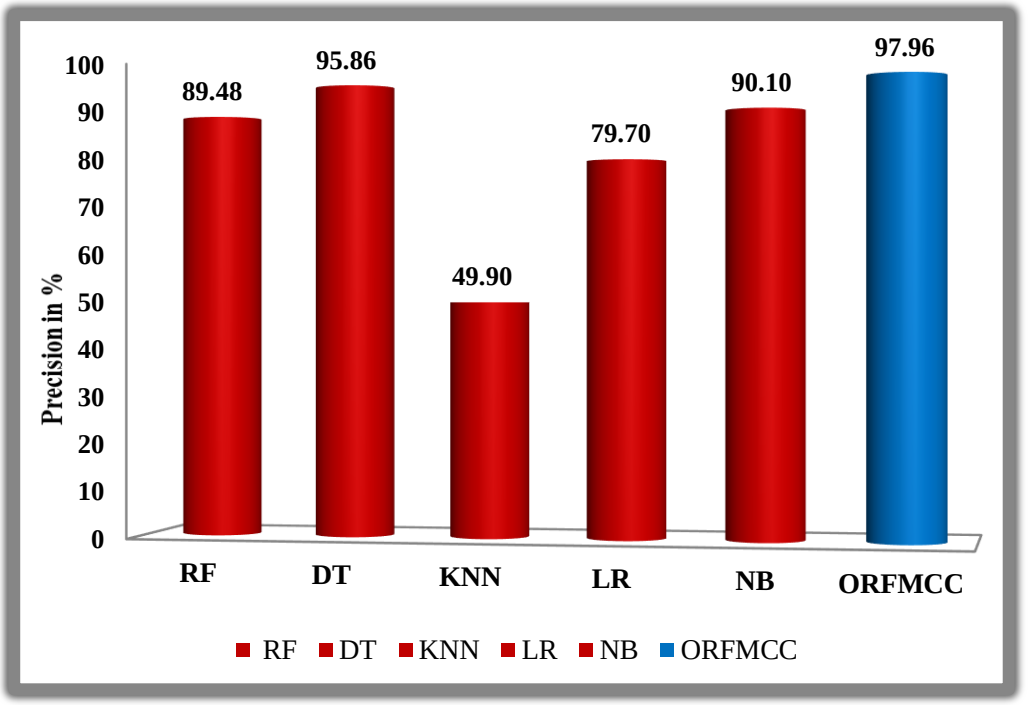

Figure 4.7 Precision

Figure 4.7 shows the Precision which represents the division of 'correctly predicted positive observations' among the 'total predicted positive observations'. The proposed ORFMCC shows better results than others.

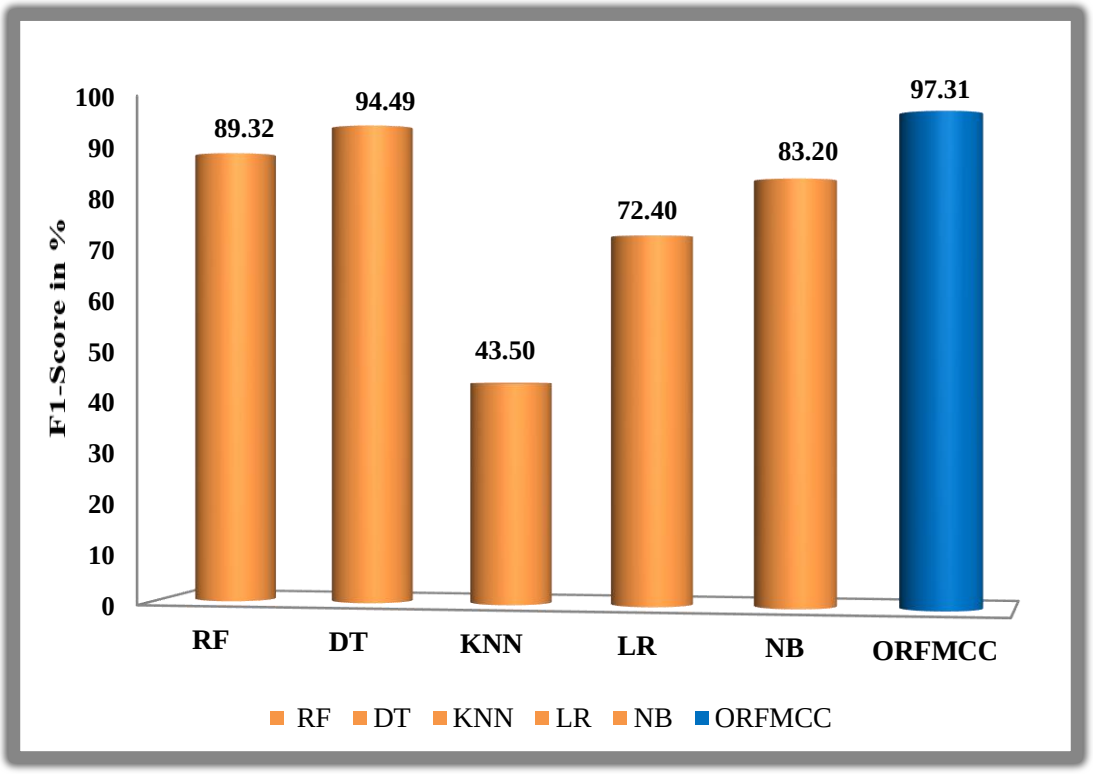

Figure 4.8 F1-Score

Figure 4.8 shows the F1-Score of the classifiers. Precision and recall were combined to calculate the F1 score of the classification performance based on the positive class predicted. The F1 score can be interpreted as the weighted average of precision and recall. Usually, F1-score is more important than the accuracy in the case of imbalanced classes. The result shows that the Optimized Random Forest is better than the other five classifiers.

Conclusion

This proposed research work focused on optimizing Random Forest for Multiclass classification (ORFMCC) to classify the students. Students' data were collected and preprocessed to optimize the Random Forest classifier. Additionally, the prediction algorithm has been created to find out the target variable. ORFMCC is developed by introducing hyperparameter tuning using cross-3 validation, and feature selection applied using feature importance score to remove irrelevant attributes. Encoding the categorical data into numeric values significantly improves the multiclass classification performance. The result produced by the proposed ORFMCC was assessed with regard to the performance metrics 'Accuracy, F1-score, Recall, and Precision'. The five-state of art classifiers Random Forest (RF), Decision Tree (DT), K- Nearest- Neighbors (KNN), Logistic Regression (LR), and Naïve Bayes (NB)' were considered for performance comparison. The proposed ORFMCC outperformed the other five classifiers. This research work recommends that researchers can apply hyper parameter tuning

and feature selection when need to improve the accuracy of the classifiers in some other data set.

References

[1] Bujang, SitiDianah Abdul, et al. "Multiclass prediction model for student grade prediction using machine learning." IEEE Access 9 (2021): 95608-95621.

[2] Tripathi, Ankita, et al. "A multi class random forest (MCRF) model for classification of small plant peptides." International Journal of Information Management Data Insights 1.2 (2021): 100029.

[3] P. Amutha, R. Priya,, Analysis of Higher Education Counselling and its awareness among Higher Secondary Students in Tamil Nadu, India, International Journal of Modern Agriculture, Volume 10, No.2, 2021.

[4] Sevastyanov, Leonid A., and Eugene Yu Shchetinin. "On methods for improving the accuracy of multi-class classification on imbalanced data." ITTMM. 2020.

[5] Hassan, Hasniza, NorBahiah Ahmad, and SyahidAnuar. "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining." Journal of Physics: Conference Series. Vol. 1529.No. 5.IOP Publishing, 2020.

[6] Thomas, Roy, and J. E. Judith. "A Novel Ensemble Method for Detecting Outliers in Categorical Data." International Journal 9.4 (2020).

[7] Thomas, Roy, and J. E. Judith. "Voting-Based Ensemble of Unsupervised Outlier Detectors." Advances in Communication Systems and Networks.Springer, Singapore, 2020.501-511.

[8] P. Amutha, R. Priya,, Conceptual Course Selection Framework for Post-Secondary Students' Enrolment in Indian Universities and Colleges, Journal of Advanced Research in Dynamical & Control Systems, Vol. 12, 03-Special Issue, 2020.

[9] Do Thi Thu Hien, Cu Thi Thu Thuy, Tran Kim Anh, Dao The Son and Cu Nguyen Giap, "Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance" International Journal of Advanced Computer Science and Applications(IJACSA), 11(11), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111135

[10] Dalton Ndirangu et al. "A Hybrid Ensemble Method for Multiclass Classification and Outlier Detection". International Journal of Sciences: Basic and Applied Research (IJSBAR) · ISSN 2307-4531, January 2019.

[11] Latha, C. Beulah Christalin, and S. CarolinJeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked 16 (2019): 100203.

[12] Basu K, Basu T, Buckmire R, Lal N, Predictive Models of Student College Commitment Decisions Using Machine Learning. *Data.*2019; 4(2):65. https://doi.org/10.3390/data4020065

[13] Mustaqeem, Anam, Syed Muhammad Anwar, and MuahammadMajid. "Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants." Computational and mathematical methods in medicine 2018 (2018).

[14] Lango, Mateusz, and Jerzy Stefanowski. "Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data." Journal of Intelligent Information Systems 50.1 (2018): 97-127.
[15] AshishDutt and MaizatulAkmarIsmail , “Can We Predict Student Learning Performance from LMS data? A Classification Approach”, Advances in Social Science, Education and Humanities Research, volume 326, 2018.
[16] Hota, Soudamini, and SudhirPathak. "KNN classifier based approach for multi-class sentiment analysis of twitter data." International Journal of Engineering & Technology 7.3 (2018): 1372-1375.
[17] Katuwal, Rakesh, and Ponnuthurai N. Suganthan. "Enhancing multi-class classification of random forest using random vector functional neural network and oblique decision surfaces." 2018 International Joint Conference on Neural Networks (IJCNN).IEEE, 2018.
[18] P. Amutha, R. Priya, A survey on educational data mining techniques in predicting student’s academic performance, International Journal of Engineering & Technology, 7 (2.33) (2018) 634-636.
[19] Wang, Ying, et al. "Using multiclass classification to automate the identification of patient safety incident reports by type and severity." BMC medical informatics and decision making 17.1 (2017): 1-12.
[20] Maryam, Noor AkhmadSetiawan, and OyasWahyunggoro. "A hybrid feature selection method using multiclass SVM for diagnosis of erythemato-squamous disease." AIP Conference Proceedings.Vol. 1867.No. 1.AIP Publishing LLC, 2017.
[21] Potdar, Kedar, Taher S. Pardawala, and Chinmay D. Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers." International journal of computer applications 175.4 (2017): 7-9.
[22] Kannan, Ramakrishnan, et al. "Outlier detection for text data." Proceedings of the 2017 siam international conference on data mining. Society for Industrial and Applied Mathematics, 2017.
[23] Chaudhary, Archana, SavitaKolhe, and Raj Kamal. "An improved random forest classifier for multi-class classification." Information Processing in Agriculture 3.4 (2016): 215-222.
[24] Farid, DewanMd, et al. "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks." Expert systems with applications 41.4 (2014): 1937-1946.