A Comparative Evaluation of Four Ensemble Machine Learning Methods for Distant Learning Classification

Abraham E. Evwiekpaefe and Biniya Ma'aruf

Department of Computer Science, Nigerian Defence Academy, Kaduna, Nigeria.

Email: aeevwikpaefe@nda.edu.ng; biniyamaaruf01@gmail.com

ABSTRACT

Predicting student performance is one of the most challenging tasks in educational organizations. The technique used to determine whether a candidate succeeds or fails in a remote learning system differs based on various criteria. Open distance learning study in this context utilized the ensemble approach to forecast the academic outcomes of students. The developed ensemble model for predicting student academic performance in an open distance learning environment used secondary data from the National Open University of Nigeria. In this experiment, four ensemble machine learning algorithms were used thus: logit boost, Random forest, Adaboost, and Bagging. Based on the data available, the bagging ensemble classifier was determined to be the best classifier, with the highest accuracy score of 98.5 percent among all ensemble and non-ensemble classifiers used in the experiment. Waikato Environment for Knowledge Analysis (Weka) software was used for the classification. However, the developed model using bagging was found to be the best for predicting student academic performance in an open distance learning environment and it can be recommended for researchers when predicting student academic performance especially in an open distance learning environment.

Keywords: Classifiers, distance learning, ensemble, machine learning, academic performance.

Reference Format:

Abraham E. Evwiekpaefe and Biniya Ma'aruf (2021), A Comparative Evaluation of Four Ensemble Machine Learning Methods for Distant Learning Classification Nigeria,
Afr. J. MIS, Vol.3, Issue 1, pp. 37 - 56.

1. INTRODUCTION

Traditional system of learning is the way of gathering and educating students at a particular place and time whereas an open distance learning creates an environment for wider ranges of cooperative and flexible opportunities devoid of place and time where individuals can participate in learning and training. These domains alleviate the delivery of large parts of education through the use of tools and materials that are easily obtained directly to the learners' at any time (Shanthini et al., 2018).

Researchers emphasized that student performance prediction is a critical component of completely customized learning environments, as well as an important component of attempts to provide excellent education. For example, higher education institutions are incorporating predictive aspects into their instructional processes to better help students (Stapel et al., 2018). Files, hard disks, CD drives, databases, and a variety of other devices may all store data. Data mining is the process of extracting relevant and essential files from a large volume of data (Tripathi & Kumar, 2019). Data mining and Knowledge Discovery in Databases have suggested a system of discovery for educational data mining (KDD). Its primary objectives are to concentrate on beneficial patterns and recognize relevant data from educational information systems such as admissions, registration, course administration, syllabus management, and other systems. These systems and initiatives work with students at many levels of education, such as schools, colleges, and universities. The primary components of effective learning are rational thinking and the interchange of viewpoints; both may be achieved through communication between students and teachers, as well as among students, throughout the learning process. However, in the educational context in general, and in e-learning in particular, pointing out the distinctive contribution of each contact has been and continues to be a topic of inquiry and debate (Stapel et al., 2018).

Furthermore, technological advancements that are utilized to improve the synergistic and media content of the web, as well as the rising quality of distribution platforms, offer a perfect setting for the growth of e-learning systems (Al-Malaise et al., 2014).Students in an open distance learning environment perform the majority of their academic activities online. In an open distance learning environment, there is little physical interaction between students and facilitators, unlike normal conventional settings, which targets to comprise better higher dimensions of flexibility and openness in terms of accessibility, syllabus (course content) or other basics of its organization.Different researchers have put a lot of effort into predicting student academic performance in an open distance learning environment; some focus on student interaction in learning management systems, while others focus on academic factors; however, research shows that there are other factors that influence student academic performance besides academic factors(Al-Malaise, et al. (2018);Afeni et al. (2019); Stapel et al. (2018)). As a result of the large number of students enrolled in the postgraduate programmes of the case study, this study concentrated on the Open distance learning students pursuing a postgraduate diploma course. This allowed for adequate data to be collected. Also, the research considered other related factors that may affect student performance, such as marital status, gender, and occupation, etc, to see if these factors will have influence the academic performance of the distance learning students. Ensemble methods are always the best in terms of performance accuracy, and other essential information is mostly provided by the model, as will be highlighted in this research work (AL-Malaise, et al., 2018).Ensemble methods have been used in diverse real world tasks like network intrusion detection, molecular bioactivity & protein locale prediction, pulmonary embolisms detection, customer relationship management, educational data mining and music recommendation. In the Knowledge Discovery in Databases- Cup (KDD-Cup) of the three years (2009-2011), all the first – place and second-place winners used ensemble methods (Dietterich 2018). Ensemble methods can also be used in almost all branches of artificial intelligence such as object detection, recognition and tracking. Therefore, the student academic performance for this research was predicted by a set of four ensemble classifiers of logit boost, Random forest, Adaboost, and Bagging.

2. REVIEW OF RELATED WORKS

The review of related literatures used in this research is shown in Table 1.

Research Gap

After conducting an extensive literature review, the paper came up with following research gaps:

The review showed that most of the works have been done in developed countries. Very few have been done in the Nigerian context that focuses on student performance prediction in distant learning environment. Another challenging issue was the choice of classification algorithm that will be used to predict student performance. Most single classifiers do not adequately predict student academic performance in terms of accuracy and minimum error rates. Ensemble methods have proved to be the best in terms of prediction. A hybrid method will therefore be applied to the student data. Previous studies hardly use this.

3. METHODOLOGY

Dataset

The dataset of National Open University of Nigeria, Faculty of Agriculture, Department of Agricultural Management Extension was collected. The dataset contains attributes from Demographic, Academic and Social factors. The dataset was cleaned and preprocessed and later divided into training and testing for model construction. The model was trained using different splitting criteria; as 60% training 40% testing, 70% training and 30% testing, 80% training and 20% testing and finally 90% testing and 10% testing and compare the result and found 80% training and 20% testing gives the best result. So, we used 80% of the dataset for training and tested with 20% of the dataset respectively. The data collected contains information such as; student id, student name, programme of study, date of birth, study center, geo-political zone, state, local government, gender, level, occupation, marital status, SSCE results, class of degree, Assignment view, Assignment submit, first semester result and second semester result respectively. These attributes are shown in Table 2.

Data Preparation and Cleaning

All of the effort involved in generating the final data set that was fed into the model is referred to as data preparation and cleaning. The data was preprocessed since it was inconsistent and noisy. To eliminate noise and inconsistencies in the data, the standard data cleaning approach was used. The attribute filter in WEKA 3.9.4 was used to eliminate information about the student's identification, such as the student's name and program. Due to the huge number of local governments (about 744), local government was also eliminated. Only information on students' prior academic success and present academic performance was kept. Since there is wide gap between the prediction classes, the attribute class balancer was used to improve the accuracy of the algorithms. After preprocessing, the performance class of the attribute was included to the dataset that was used to predict student performance. EXCELLENT, GOOD, and POOR were the three performance levels. To aid in the reduction of the number of classes in the confusion matrix, as a large number of classes in the confusion matrix might cause errors in some classification methods. The classes are as follows: if a student's first and second semester results are both above or equal to 4.00, the model will classify him as excellent; if the first and second semester results are both above 2.50 and below 3.99, the model will classify him as good; and if the first and second semester results are both below 2.50, the model will classify him as poor.

Constructing the Model

Four machine learning methods are independently implemented in the development of the student performance model. However, using the average combination rule, an

ensemble bagging technique was developed with Decision tree, K-nearest neighbor, Multilayer Perception, and Random forest as the base classifier, comparing the performance of the individual machine learning algorithms with the model developed using Bagging Meta ensemble technique. The model was built using the WEKA open source program 3.9.4.

Model Design

i. Input dataset of students
ii. Apply preprocessing to the dataset
iii. Perform cross-validation
iv. Apply machine learning algorithms to the dataset independently
v. Develop classification model
   a) Find classification hypothesis
   b) Perform voting process by combining classifiers using posterior probability rules
vi. Compare the machine learning algorithms with the hybrid classification model

Figure 3.1 shows the architecture of the methodology design for this work

Model Performance Evaluation

With the aid of a confusion matrix, the performance of the predictive model is assessed using evaluation metrics. The matrix is N × N, where N is the number of target values (classes), and it represents the number of right and wrong predictions produced by the model in comparison to the actual data outcomes. The confusion matrix data is used to assess the models' performance. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are the elements in the confusion matrix that are used to calculate the Accuracy, Precision, Recall, and F-measure of each predicted and actual class in the proposed student's performance prediction model. However, model accuracy was not the only factor examined; additional assessment metrics such as root mean square error (RMSE), True positive rate (TPR), and True negative rate (TNR) were also taken into account. A confusion matrix is depicted in Table 3.

True Positive (TP) represents the number of instances that were actually positive and were predicted to be positive, False Positive (FP) represents the number of instances that were actually negative but were predicted to be positive, and False Negative (FN) represents the number of instances that were actually positive but were predicted to be negative, according to Table 3. The following is a breakdown of the performance indicators:

Accuracy: The total number of cases correctly categorized is referred to as accuracy. It is one of the assessment measures that researchers may use to determine the efficiency of any machine learning system. It's worked out using the formula 1.

A= (TP+TN)/(TP+TN+FP+FN) ------------------- (1)

where:

A is the accuracy.

P is the positives

N is the negatives

TP: True Positive which is the number of positive instances that are correctly classified.

TN: True Negative is the number of negative instances correctly classified.

FN: False Negative is the number of negative instances incorrectly classified.

FP: False Positive is the number of positive instances incorrectly classified.

Precision: The sum of the number of correct positive classifications divided by the total number of positive classifications is known as sensitivity. It can also be calculated using equation 2.

P=TP/(TP+FP) -------------------------- (2)

where:

P is the precision

TP: True Positive which is the number positive instances that are correctly classified.

FP: False Positive is the number of negative instances incorrectly classified.

Recall: Specificity is the number of right classifications divided by the total number of positives, which is referred to as recall. The formula 3 is used to assess it.

R=TP/(TP+FN) =TP/P ----------------------- (3)

where:

R is the recall

P is the precision

TP: True Positive which is the number of positive instances correctly classified

FN: False Negative which is the number of negative instances incorrectly classified

F-Score: Is a harmonic mean of recall and accuracy. The F-score is calculated using formula 4.

F=2PR/(P+R) ----------------------------- (4)

W

where:

F is the F-score

P is the precision

R is the recall

4. RESULTS

The visualization results of some of the attributes

Geo-political Zones

The number of geo-political zones in Nigeria is depicted in figure 4.3. In each geo-political zones it gives us information of the number of student that study Agricultural Management Extension at National Open University of Nigeria and those that their result are predicted as excellent, good or poor. From the Figure 2, it was found that the majority of student that study Agricultural Management Extension at NOUN are from North-central followed by south-west which is the second then, south-south then, north-west then, south-east and lastly north-east.

States

The Figure 3 displays the number of students from each state across the country. Students from Ondo state had the highest number of population at the Faculty of Agriculture. The result depicts the states with the number of students that have excellent, good and poor in their results.

Occupation

Figure 4 gives us information on the student that are working, the data is classified as working and not working. Those that are civil servant and business are classified as working and those that are students are classified as not working. Money constraints is one of the major draw back for srudents failure most especially in examinations since it play a major roll in acquiring learning materials. The figure 4 shows that majority of students do well in school when they have something to do.

Gender

Figure 5 gives the statistics of the gender of the students. The results obtained clearly depicts that majority of people that study Agricultural Management Extension at National Open University of Nigeria are males which can be generalized that males pursue western education more than females in the department.

Marital Status

Marital status is one of the major attribute that have effect in the student academic performance in schools. The result showed that out of 755 student that study Agricultural Management Extension at National Open University of Nigeria, 380 students are single where 352 are married and only 23 are divorced as shown in Figure 6.

Performance Class

Figure 7 depicted their final result for the first and second semester. The final result is classified as excellent, good and poor. Students with their average CGPA score ranking above 4.00 are classified as excellent those thus with CGPA from 2.50 to 3.99 is classified as good, and the students with CGPA below 2.50 were classified as poor.

Ensemble Classifiers Without Feature Selection

LogitBoost Classifier: LogitBoost was one of the Meta classifiers utilized in this experiment. The classifier was validated using 10-fold cross validation on the students dataset acquired from NOUN, which had 755 occurrences and 15 instances. As demonstrated in Figure 8, the classifier achieved 96.56 percent accuracy without feature selection.

Adaboost: It is a meta estimator that starts by fitting a classifier to the original dataset, then fits further copies of the classifier to the same dataset, but adjusts the weights of wrongly classified instances such that future classifiers focus more on difficult cases. The accuracy obtained without feature selection is95.50%. See Figure 9.

Random Forest: Random forest is another ensemble classifier that, in most cases, produces accurate predictions. It was also utilized in this experiment using the same dataset as th method above. As illustrated in Figure 10, the random forest method has an accuracy of 91.39 percent, with 690 properly categorized cases and 65 erroneously classified examples.

Bagging (Our Approach): The Bagging Meta classifier was chosen since it is one of the best classifiers for predicting student achievement. The goal was to change the internal parameter of the created model in order to improve the classifier's performance accuracy. The Bagging classifier was modified to create a model in which the basis classifiers were Decision tree, K-nearest neighbor, Multilayer Perception, and Random forest. The model was trained by utilizing the average combination rule to take the vote of their forecasts. There were 15 characteristics and 755 occurrences in the data. The model is constructed by utilizing 10-fold cross validation to combine the votes of the different classifiers. After the data was separated, 80 percent of it was utilized to train the model, while the remaining 20% was used to test it. Ten iterations were used to train the model. The quantity of data utilized for decreased error-pruning was determined using only three folds. With the data, the meta bagging classifier was able to achieve a performance accuracy of 93.64 percent, with 707 properly categorized examples and a low error rate of 48 wrongly classified instances. As illustrated in Figure 11, the model also had a mean absolute error of 0.0923, a root mean square error of 0.1811, a relative absolute error of 44.5068, and a root relative square error of 56.5674.

Comparison of Accuracy With Ensemble Classifiers

Here is the comparison of the results with other ensemble classifiers without feature selection as shown in Table 4.

Models Generated from Individual Classifiers with Feature Selection and Class Balancing

Models constructed using ensemble classifiers presented the greater results after conducting feature selection and class balancing, as illustrated in Figures 12 through Figure 15. Among the ensemble classifiers, the bagging approach adopted had the highest performance accuracy of 98.4946 percent, indicating that the bagging meta classifier is one of the best classifiers for feature selection and class balance.

Comparison of Classifiers with Feature Selection and Class Balancing

Table 5 showsthe comparison of the accuracies of the ensemble classifiers with feature selection. However, ensemble bagging classifier is recommended because it gave the highest performance accuracy of 98.5 percent with feature selection and class balancing. See Table 5.

5. DISCUSSION

Though existing models took into account aspects such as a student's familial history, social characteristics, and academic indicators, key elements such as profession, a student's past academic record which has a substantial impact and so on are often overlooked. In an open distance learning setting most especially postgraduate diploma, students have only one chance of either pass or fail because of the duration of the programme for PGD students. The ensemble model is ideal for investigating all of the elements that go into creating a strong and trustworthy model. The ensemble model's outputs allow for the efficient and accurate prediction of student performance, as well as the identification of students who are at danger of failing or dropping out.

Previous academic records have a major impact on present performance, as do other criteria such as marital status, assignment submission, and semester results, all of which play a part in a student's overall success. Classification techniques were used to create the ensemble model. After preprocessing the dataset, and removing features that were not required for the experiment, the ensemble model was created by the taking the vote of their predictions and combined in an ensemble using average combination rules. Because there is such a big disparity between the three classes, the attribute class balancer was used to enhance the model's performance. Tables 5 illustrated the results of the proposed model after class balancing for ensemble classifiers. Before class balancing, Logitboost had the greatest accuracy of 96.56 percent, but after class balancing, bagging had an overall accuracy of 98.50 percent with the fewest root mean square errors of 0.1115.

With the available data set, the model was able to reach 98.50 percent accuracy and identify the key determinants impacting academic achievement. The root mean square error (RMSE) was also taken into account in this study, which clearly shows that the bagging ensemble classifier got the lowest root mean square error of 0.1115, as shown in Figure 15. The research shows that the bagging meta classifier is best suited for this type of study based on the results gained from the models built.

6. CONCLUSION

Bagging Meta classifier has the capability of improving the performance accuracy and efficiency of student's academic performance prediction models. The research evaluated the performance of four (4) ensemble machine learning algorithms in developing the student academic performance model. The result showed that the developed bagging method outperformed the other meta ensemble classification models.

REFERENCES

1. Abdulazeez, Y., &Abdulwahab, L. (2018). Application of classifications mode to predictstudents' academic performance using classifiers ensemble and synthetic minority over sampling techniques. *Bayero Journal of Pure and Applied Sciences. 11 (2), 142-148.*

2. Abubakar, Y., & Ahmad, N.B. (2017). Prediction of students' performance in E-learning environment using Random Forest. *International Journal of InnovativeComputing. 7(2), 1-5.*

3. Adebayo, A. O., &Chaucer, M.S. (2019). Data mining classification techniques on theanalysis of student performance, *Global Science Journal. 7(4), 80-84.*

4. Afeni, B. O., Oloyede, I. A. & Okurinboye, D. (2019). Students performance predictionusing classification Algorithms. *Journal of Advances in Mathematics and Computer Science. 30(2), 1-9.*

5. Al-Malaise, A., Malibari, A. & Alkhozae, M. (2014). Student performance predictionsystem using multi agent data mining technique. *International Journal of Data Mining and Knowledge Management Process. 4(5), 1-20.*

6. Almasri, A., Celebi. E. & Alkhawaled, R.S. (2019) EMT: Ensemble meta-based tree model for predicting student performance. *Hindawi Scientific Programming. 9(13), 2-6.*

7. Dietterich, T.G. (2018) Ensemble Methods in Machine Learning. *International Journal Data Mining and Knowledge Management Process, 5(1), 576-584.*

8. Hussain, S., Dahan, N. A., Ba-Alwab, F. M. & Ribata, N. (2018). Educational data mining of student's academic performance using WEKA. *Indonesian Journal of Electrical Engineering & Computer Science,9(2), 447-459.*

9. Jamil, A., Ahsan, M., Farooq, T., Hussain, A. & Ashraf, R. (2018). Student performanceprediction using algorithm of data mining, *4th International Conference onComputing, Engineering and Design.* Singapore, India, *9(5),246-248.*

10. Kaur, P., Singh, M. & Josan, G. S. (2015). Classification and prediction-based datamining algorithm to predict slow learners in education sector,*3rd International Conference on Recent Trends in Computing. 57, 500-508.*

11. Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019). Comparative study of supervised learning algorithms for student performance prediction. *Journal of ICAIIC. 1(2), 124-127. https://doi.org/10.1109/5386-7822*

12. Rawat, K. S.&Malhan, I.V. (2018). A hybrid classification method based on machinelearningClassifiers to predict performance in educational data mining, *proceeding of 2nd International Conference on Communication, Computing and Networking*, Lecture notes in Network and System, Mumbai, India. 35(2), *46-47.*

13. Shanthini, A., Vinidhini, G., & Chandrasekaran, R.M. (2018). Predicting studentacademic

performance using meta decision tree classifier. *Journal of Computer Science.14(5), 654-662.*

14. Stapel, M., Zheng, Z., & Pinkwart, N. (2018). An Ensemble Methods to Predict StudentPerformance in an Online Math Learning Environment.*9th International Conference On Educational Data Mining.* Istanbul, Turkey. 23(2), *237-240.*

15. Tripathi, A. & Kumar, A. (2019). Analysis of Educational Data Mining Techniques.*InternationalJournal of Computer Science and Mobile Computing, 8(1), 8-15.*

16. Usman, U.I., Aliyu, S., Ahmad, B.I. & Yusuf, A. (2019) A comparative study of base classifiers in predicting students performance based on interaction with LMSplatform. *FUDMA Journal of Science. 3(1), 231-239.*

Table 1: Review of Related Works

Author's Name	Description	Outcomes
Kaur et al. (2015)	Proposed a technique for prediction on the dataset of 152 students'. To predict and analyze student's performance as well as slow learners among them.	The researchers used five classifiers as multilayer perception, Naïve Bayes, SMO, J48 and REPTree. After conducting the experiment, it was found that multilayer perception out performs the others with 75% accuracy.
Abubakar and Ahmad (2017)	Predict student performance in e-learning environment using random forest. A dataset of 26 student record was fed into MATLAB software using 10-fold cross validation. Random forest, Naïve Bayes and K-nearest neighbor was used as the classification algorithms.	After the experiment was conducted, Naïve Bayes outperforms the rest with the highest accuracy and RMSE but Random forest produced more information than NB.
Abdulazeez and Abdulwahab (2018)	Investigated the application of classification models in predicting student academic performance using synthetic minority over sampling technique. Dataset of Federal University Dutse consisting of 206 student records was used. WEKA 3.9.1 was used as the tool. Stacking classifier was developed using J48, KNN, and SMO	The experiment was able to reveal that stacking ensemble classifier works better in the context of their research with an accuracy of 96.7%
Dietterich (2018)	Briefly investigated methods for constructing ensembles.	The results show that ensemble classifiers are always better than single classifier.
Hussain et al. (2018)	Used a data of student collected from Assam, India to evaluate student academic performance. The researcher used Random Forest, PART, J48 and BayesNet to carry out the experiment.	The researchers used 300 instances and 24 attributes to perform the experiment. After conducting the experiment, Random Forest outperforms the rest with

		performance accuracy of 84%.
Jamil et al. (2018)	The researchers collected data through questionnaire and survey methods. 257 samples from students of Aj-jar aha were obtained.	12 attributes were used among the five classifiers; J48, REPTree, ZeroR, Naïve Bayes and Random Forest. Out of the five classifiers, Random forest gives the best performance accuracy of 92%
Rawat and Malhan (2018)	Proposed a hybrid method using four machine learning algorithms; decision tree, Naive Bayes, K-nearest neighbor and multilayer perception.	To obtain the efficiency of accuracy, all the four classifiers are grouped using voting approach. A 10-fold cross-validation is used to predict the evaluation accuracy. Hence, the proposed method proved to achieve the highest accuracy
Shanthini et al. (2018)	Introduced a model using ensemble method for predicting student performance using the marks obtained from 10 subjects for 2 semesters.	After carrying out the experiment, it was found that Boosting obtained the highest performance accuracy of 97%.
Stapel et al. (2018)	Proposed a new approach that combines classifiers in an ensemble to predict the performance of online math students, the data for the entire year of 2015 was used for the analysis.	Based on the analysis, it was found that from the 12 cases the three scope classifiers agreed to have contributed the wrong class of student, thereby, the ensemble cannot predict the students correctly. But, from all the classifiers that was used, the proposed ensemble method shows the best prediction accuracy of 73.5%.
Adebayo and Chaubey (2019)	Proposed a model called KNIME used for classification and also performing other analysis such as clustering on samples of dataset.	KNIME has proved to be one of the best tools for classification that calculates the student's relationship for those on the greater performance and those on the lower performance.
Almasriet al. (2019)	Proposed a model called Ensemble Meta Decision Tree (EMT) that combined two classification techniques.	The model is built along with an exhaustive investigation of the performance on the academic student training set. The performance of the proposed model proved that predictions of ensemble methods are better than that of single classifier with accuracy of 98%.

Mohammadi et al. (2019)	Used K-nearest neighbor, Decision tree, and Naïve Bayes classifiers for the dataset of 230 students from Kabul University.	Having conducting the analysis, K-nearest neighbor gives the best performance accuracy of 0.5464% followed by Decision tree with the performance accuracy of 0.5325%.
Usman et al. (2019)	Conducted a comparative study of based classifiers to predict student performance based on interaction with LMS platform	The result obtained indicates that Decision tree is the best with 84.1% accuracy which outperforms NB and KNN.

Table 2: Attributes Selection Table

Attributes	Data types	Description
1. Student ID	Numeric	(40001251-192016730)
2. Study Center	String	(Abakaliki,….Wukari study center)
3. Geo-zones	String	(North west,….South-south)
4. States	String	(Abia,…..Zamfara)
5. Occupation	String	(Yes, No)
6. Gender	String	(Male, Female)
7. Marital Status	String	(Married, Single, Divorced)
8. English	String	(A1,…,F9)
9. Mathematics	String	(A1,…,F9)
10. Entry Result	String	(First class,..Pass. Distinction,...Pass)
11. First Semester Result	Numeric	(0.00,…,5.00)
12. Second Semester Result	Numeric	(0.00,…,5.00)
13. Assign View	Numeric	(1-6)
14. Assign Submit	Numeric	(1-6)
15. Performance	String	(Excellent, Good, Poor)

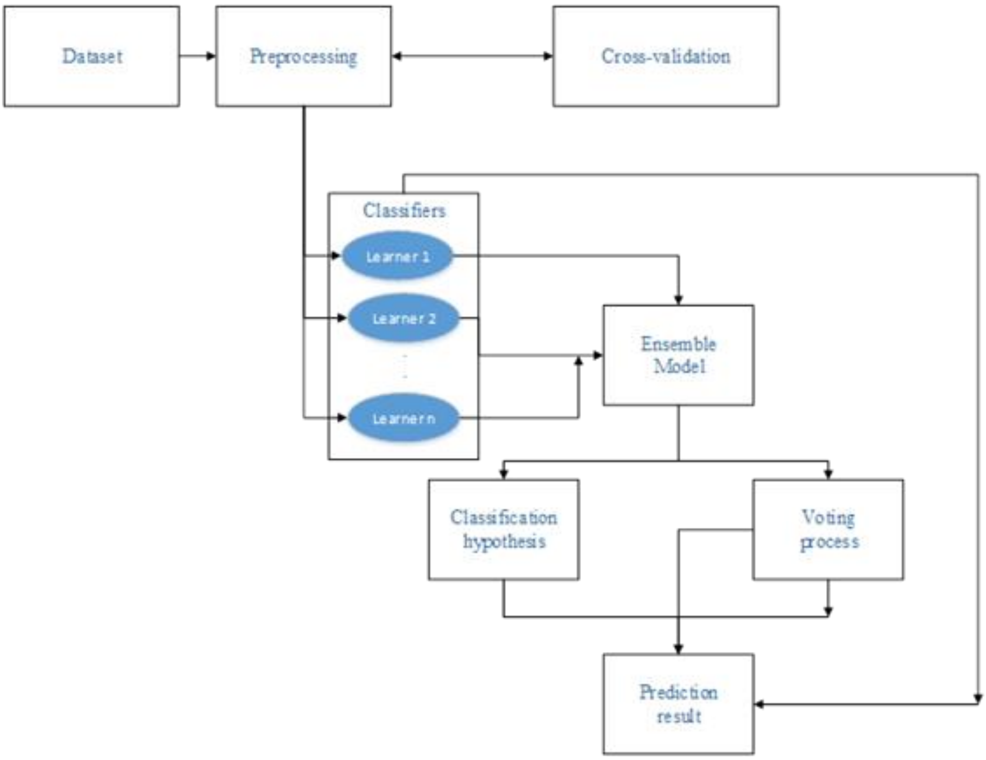


Figure 3.1: Architecture of Methodology Design

Table 3: Confusion Matrix

		PREDICTED CLASS	
		POSITIVE	NEGATIVE
ACTUALCLASS	POSITIVE	TP	FP
	NEGATIVE	FN	TN

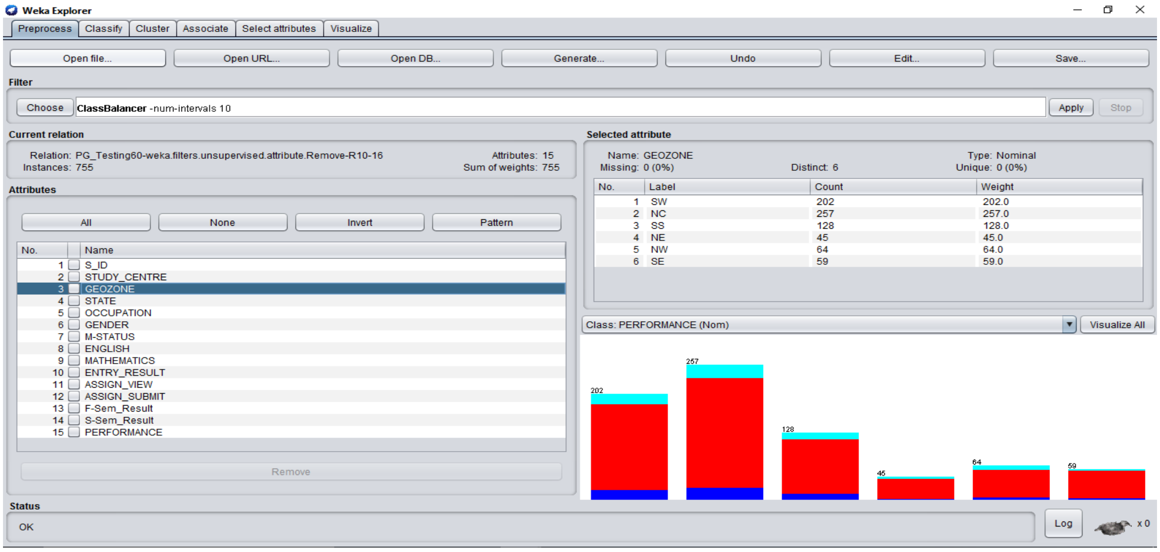


Figure 2: Geo-political Zones

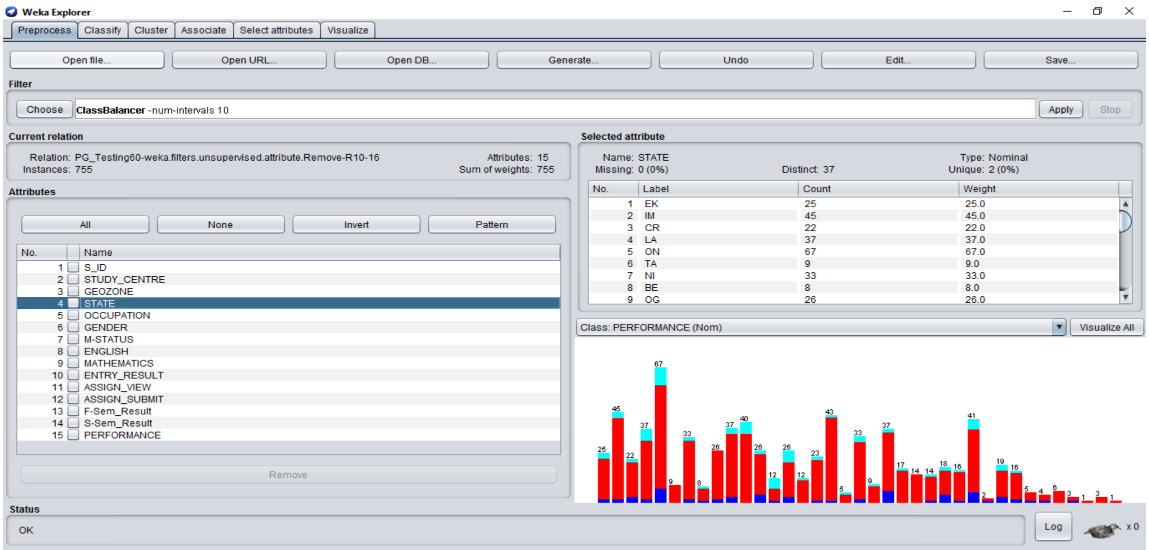


Figure 3: States

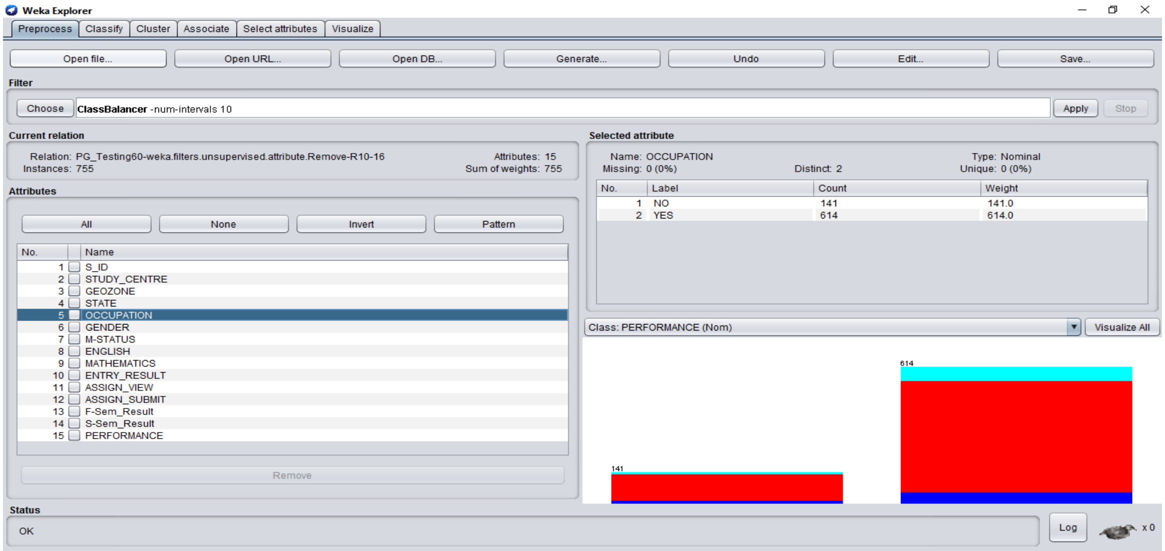


Figure 4: Occupation

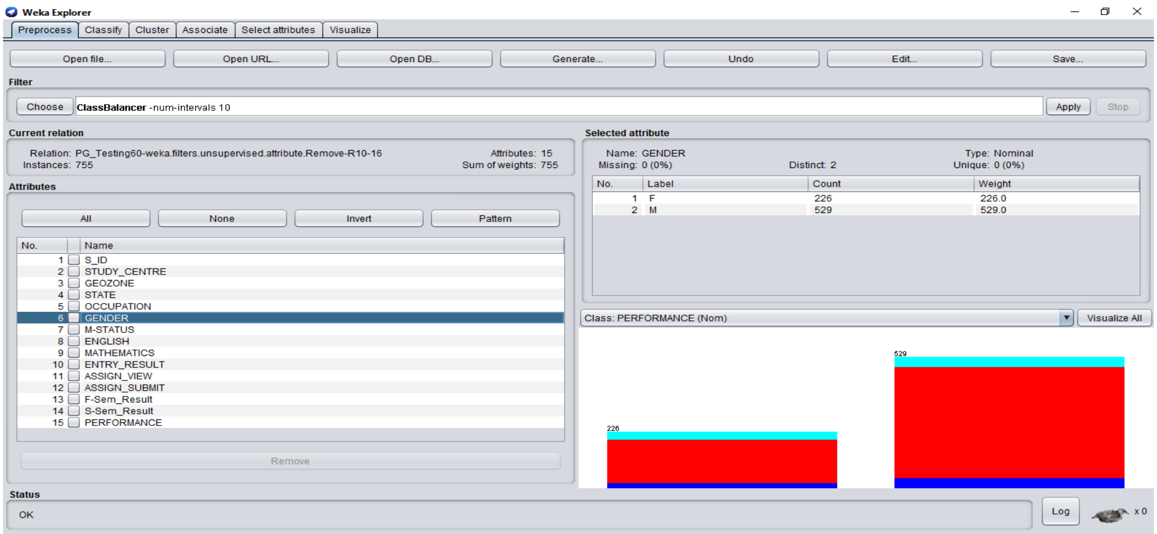


Figure 5: Gender

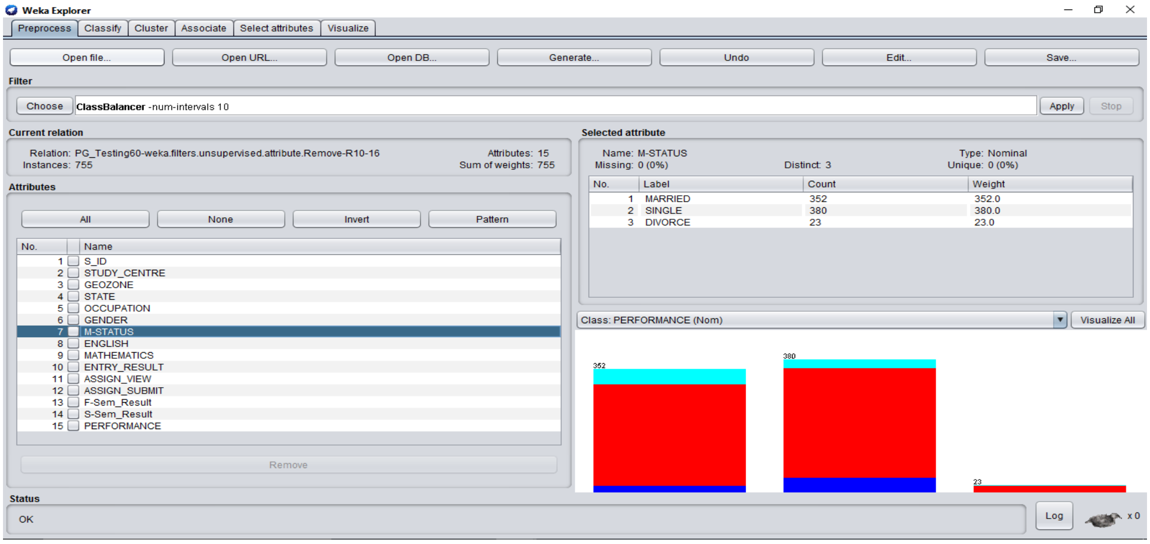


Figure 6: Marital Status

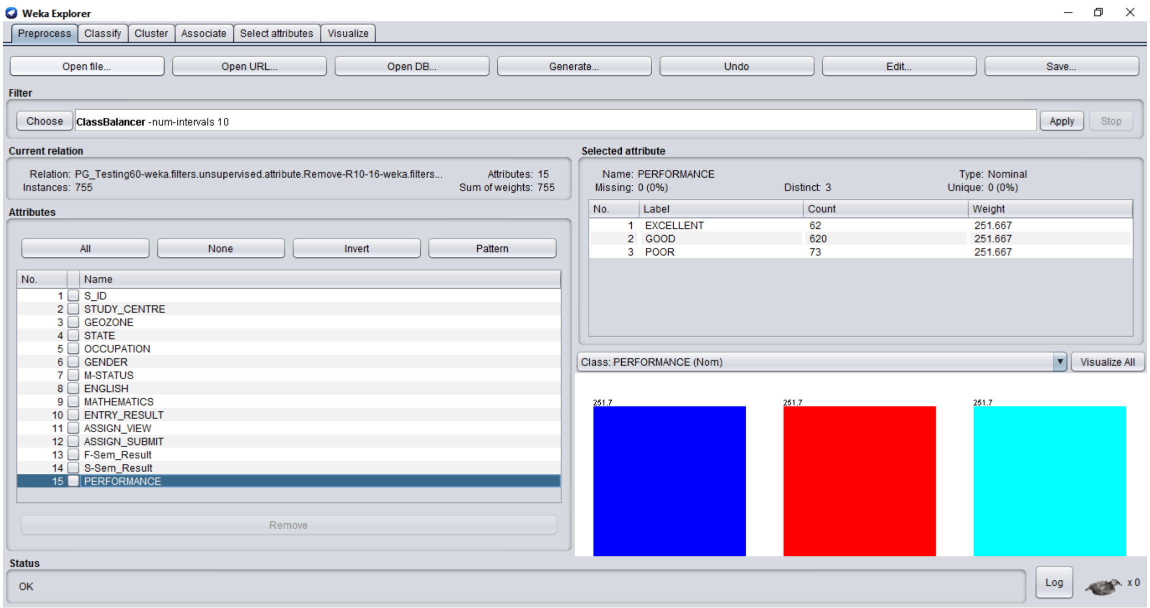


Figure 7: Predicted Classes.

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         729               96.5563 %
Incorrectly Classified Instances        26                3.4437 %
Kappa statistic                          0.888
Mean absolute error                      0.0342
Root mean squared error                  0.1288
Relative absolute error                 16.4915 %
Root relative squared error             40.1053 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.887    0.006    0.932      0.887    0.909      0.902    0.998     0.971     EXCELLENT
                 0.981    0.104    0.977      0.981    0.979      0.882    0.994     0.999     GOOD
                 0.904    0.012    0.892      0.904    0.898      0.887    0.997     0.972     POOR
Weighted Avg.    0.966    0.087    0.965      0.966    0.965      0.884    0.994     0.994

=== Confusion Matrix ===

   a   b   c   <-- classified as
  55   7   0 |   a = EXCELLENT
   4 608   8 |   b = GOOD
   0   7  66 |   c = POOR
```

Figure 8: LogitBoost Classifier Result without Feature Selection

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         721               95.4967 %
Incorrectly Classified Instances        34                4.5033 %
Kappa statistic                          0.8517
Mean absolute error                      0.0375
Root mean squared error                  0.1586
Relative absolute error                 18.0821 %
Root relative squared error             49.3839 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.855    0.010    0.883      0.855    0.869      0.857    0.986     0.888     EXCELLENT
                 0.977    0.148    0.968      0.977    0.973      0.844    0.977     0.994     GOOD
                 0.849    0.010    0.899      0.849    0.873      0.861    0.986     0.920     POOR
Weighted Avg.    0.955    0.123    0.954      0.955    0.955      0.847    0.979     0.978

=== Confusion Matrix ===

   a   b   c   <-- classified as
  53   9   0 |   a = EXCELLENT
   7 606   7 |   b = GOOD
   0  11  62 |   c = POOR
```

Figure 9: Adaboost Classifier without Feature Selection

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         690               91.3907 %
Incorrectly Classified Instances        65                8.6093 %
Kappa statistic                          0.6561
Mean absolute error                      0.1108
Root mean squared error                  0.2007
Relative absolute error                 53.4085 %
Root relative squared error             62.4777 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.419    0.001    0.963      0.419    0.584      0.618    0.994     0.915     EXCELLENT
                 0.997    0.467    0.907      0.997    0.950      0.683    0.987     0.997     GOOD
                 0.630    0.001    0.979      0.630    0.767      0.769    0.995     0.957     POOR
Weighted Avg.    0.914    0.383    0.919      0.914    0.902      0.686    0.988     0.986

=== Confusion Matrix ===

   a   b   c   <-- classified as
  26  36   0 |   a = EXCELLENT
   1 618   1 |   b = GOOD
   0  27  46 |   c = POOR
```

Figure 10: Random Forest Classifier Result without Feature Selection

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         707               93.6424 %
Incorrectly Classified Instances        48                6.3576 %
Kappa statistic                          0.7617
Mean absolute error                      0.0923
Root mean squared error                  0.1811
Relative absolute error                 44.5068 %
Root relative squared error             56.3674 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.001    0.969      0.500    0.660      0.679    0.997     0.961     EXCELLENT
                 0.997    0.341    0.931      0.997    0.963      0.772    0.993     0.999     GOOD
                 0.795    0.001    0.983      0.795    0.879      0.873    0.997     0.969     POOR
Weighted Avg.    0.936    0.280    0.939      0.936    0.930      0.774    0.994     0.993

=== Confusion Matrix ===

   a   b   c   <-- classified as
  31  31   0 |   a = EXCELLENT
   1 618   1 |   b = GOOD
   0  15  58 |   c = POOR
```

Figure 11: Bagging Meta Classifier Model Generated without Feature Selection

Table 4: Comparison of Accuracy with Ensemble Classifiers Result without Feature Selection

Classifier	Accuracy (%)	Precision	Recall	F-measure	RMSE

LogitBoost	96.56%	0.965	0.966	0.965	0.1288
Random forest	91.39%	0.919	0.914	0.902	0.2007
Adaboost	95.50%	0.954	0.955	0.955	0.1586
Our approach (Bagging)	**93.64%**	**0.939**	**0.936**	**0.930**	**0.1811**

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         734.91             97.3391 %
Incorrectly Classified Instances        20.09              2.6609 %
Kappa statistic                          0.9601
Mean absolute error                      0.0364
Root mean squared error                  0.1219
Relative absolute error                  8.1794 %
Root relative squared error             25.8535 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.968    0.006    0.988      0.968    0.978      0.967    0.999     0.997     EXCELLENT
                 0.966    0.023    0.955      0.966    0.960      0.940    0.995     0.992     GOOD
                 0.986    0.011    0.978      0.986    0.982      0.973    0.999     0.998     POOR
Weighted Avg.    0.973    0.013    0.974      0.973    0.973      0.960    0.997     0.996

=== Confusion Matrix ===

   a      b      c     <-- classified as
 243.55   8.12   0    |      a = EXCELLENT
   2.84 243.14   5.68 |      b = GOOD
   0      3.45 248.22 |      c = POOR
```

Figure 12: LogitBoost Classifier Model Generated with Feature Selection

Classifier output

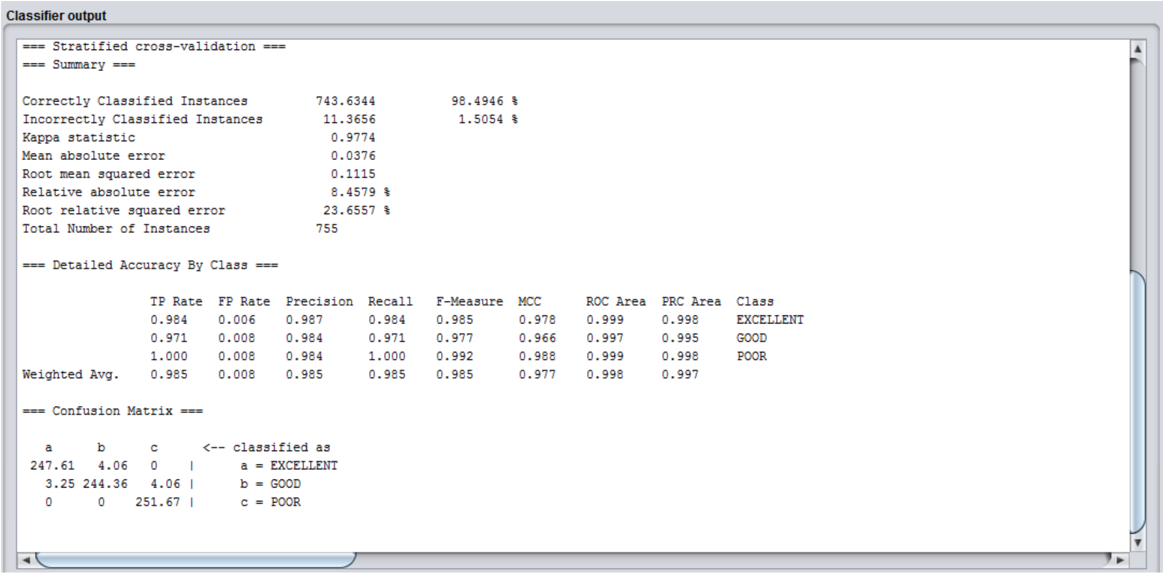
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         732.4746         97.0165 %
Incorrectly Classified Instances        22.5254          2.9835 %
Kappa statistic                          0.9552
Mean absolute error                      0.0414
Root mean squared error                  0.1329
Relative absolute error                  9.3118 %
Root relative squared error             28.1873 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.952    0.006    0.987      0.952    0.969      0.954    0.999     0.998     EXCELLENT
                 0.973    0.031    0.940      0.973    0.956      0.934    0.996     0.993     GOOD
                 0.986    0.007    0.985      0.986    0.986      0.979    0.999     0.998     POOR
Weighted Avg.    0.970    0.015    0.971      0.970    0.970      0.956    0.998     0.996

=== Confusion Matrix ===

   a      b      c      <-- classified as
 239.49  12.18   0    |      a = EXCELLENT
   3.25 244.77   3.65 |      b = GOOD
   0      3.45 248.22 |      c = POOR
```

Figure 13: Random Forest Classifier Model Generated with Feature Selection

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         703.8604         93.2265 %
Incorrectly Classified Instances        51.1396          6.7735 %
Kappa statistic                          0.8984
Mean absolute error                      0.1177
Root mean squared error                  0.2046
Relative absolute error                 26.4722 %
Root relative squared error             43.3949 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.022    0.956      0.935    0.945      0.919    0.992     0.977     EXCELLENT
                 0.889    0.046    0.906      0.889    0.897      0.847    0.963     0.943     GOOD
                 0.973    0.034    0.935      0.973    0.953      0.930    0.990     0.966     POOR
Weighted Avg.    0.932    0.034    0.932      0.932    0.932      0.898    0.981     0.962

=== Confusion Matrix ===

   a      b      c      <-- classified as
 235.43  16.24   0    |      a = EXCELLENT
  10.96 223.66  17.05 |      b = GOOD
   0      6.89 244.77 |      c = POOR
```

Figure 14: Adaboost Classifier Model Generated with Feature Selection



Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         743.6344          98.4946 %
Incorrectly Classified Instances        11.3656           1.5054 %
Kappa statistic                          0.9774
Mean absolute error                      0.0376
Root mean squared error                  0.1115
Relative absolute error                  8.4579 %
Root relative squared error             23.6557 %
Total Number of Instances              755

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.006    0.987      0.984    0.985      0.978    0.999     0.998     EXCELLENT
                 0.971    0.008    0.984      0.971    0.977      0.966    0.997     0.995     GOOD
                 1.000    0.008    0.984      1.000    0.992      0.988    0.999     0.998     POOR
Weighted Avg.    0.985    0.008    0.985      0.985    0.985      0.977    0.998     0.997

=== Confusion Matrix ===

   a      b      c     <-- classified as
 247.61   4.06   0    |      a = EXCELLENT
   3.25 244.36   4.06 |      b = GOOD
   0      0    251.67 |      c = POOR
```

Figure 15: Bagging Meta Classifier Model Generated with Feature Selection

Table 5: Comparison of Accuracy of Ensemble Classifiers with Feature Selection and Class Balancing

Classifier	Accuracy	Precision	Recall	F-Measure	RMSE
LogitBoost	97.34%	0.974	0.973	0.973	0.1219
Random Forest	97.02%	0.971	0.970	0.970	0.1329
Adaboost	93.23%	0.932	0.932	0.932	0.2064
Bagging	**98.50%**	**0.985**	**0.985**	**0.985**	**0.1115**