Ensemble Voting Machine Learning Model for Prediction of Campus Placement of the Student

B. Kalaiselvi[1, a)] and S. Geetha[2, b)]

[1]*Department of Computer Science, Government Arts College, Udumalpet, Tamil nadu, India.*
[2]*Department of Computer Science, Government Arts and Science College for Women, Puliakulam, Coimbatore, Tamil Nadu, India.*

[a)] Corresponding author: bkalaimca@gmail.com
[b)] geet_shan@yahoo.com

Abstract. Placements in campus interviews are the dream of every student in college. Placement in campus interviews is a vital measure of an educational institution's standards and student performance. Machine learning with the knowledge discovery process helps to forecast the student's performance in on-campus interviews. This paper suggested an ensemble model-based voting classifier with BayesNet and J48 is used to classify the student's academic data and forecast the placement opportunity. This work compares two ensemble stacking models and a voting-based classification model with J48 for obtaining an efficient model of placement prediction. Both ensemble stacking models use BayesNet and J48 classifiers as the base classifiers. The J48 classifier is used as the meta classifier in one stacking process and the voted perceptron is used in another. In the ensemble voting model, BayesNet and J48 are used as the base classifiers and the probability average of a class of base classifiers is used for the combination rule. The ensemble voting model gains high accuracy with a minimum error rate than other models. This model produced 91% of accuracy in the placement prediction. J48 and BayesNet classifiers are combined with probability average-based combination rules in the ensemble voting model.

INTRODUCTION

Indian higher education institutions provide the highest count of placements in the world. At the same time, the count of higher educational institutions is also very high. So, each and every institution needs to improve the quality of education and increase placement opportunities. Educational Data Mining (EDM) plays a vital role in improvising the quality of education, identification of learning difficulties, forecasting weak students and placement opportunities and updating educational settings. EDM creates a knowledge base for institutions and students by extracting hidden knowledge from educational data. Obtaining good scores in semester exams and getting a placement with a high package from premier organizations by the student are the most used evaluation pattern in college education. Every higher education institution needs to forecast placement opportunities for the students in campus interviews at an earlier stage. It facilitates tuning the students and improving their performance in campus placement interviews. In the forecasting of placement status, a high accuracy rate is necessary to group the students by the placement cell to provide more training and attention, so a better system is required for placement status forecasts with a high accuracy rate.

This paper presents an ensemble machine-learning classification model for the placement status forecasting process. Combined classification techniques are used in stacking and voting models to forecast the placement details. The performance of this ensemble model in the student data is examined and the best ensemble model for the further placement status forecasting process of the student data is selected.

International Conference on Innovations in Robotics, Intelligent Automation and Control
AIP Conf. Proc. 2914, 050022-1–050022-6; https://doi.org/10.1063/5.0175866
Published by AIP Publishing. 978-0-7354-4770-7/$30.00

LITERATURE SURVEY

S Dutta et al., [1] proposed an ensemble voting classifier for the campus placement forecasting process. Gradient boosting and Extra Tree classifiers are combined in this ensemble voting technique. In this process, existing data set from Kaggle is used and predicts the placement status with 86.05% accuracy.

B Sen et al., [2] analyzed placement test marks for the prediction process with C5, SVM, ANN, and Logistic regression. The C5-based decision tree algorithm produced high accuracy (95%) than the other three models. The author identified scholarship, GPA and previous test scores are important factors in the prediction of placement test marks.

Dech Thammasiri et al., [3] includes SMOTE for class imbalance with SVM to obtain the best accuracy in student attrition forecast and identified vital attributes for accuracy in prediction. 90.24% accuracy was obtained by this model and it's better than SVM, NN, LR, and decision tree classifiers.

V K Harihar [4] compared MLP with Tree-based logistic model and SVM classifiers in the UG students' placement possibilities. Different datasets and measures are used in the prediction process. RMSE, Accuracy, F1Score, and ROC performance metrics are used to analyze the classifier's performance. The tree-based logistic model predicts the possibilities of the placement for the students with 99.5% of accuracy.

Shreyas. H et. al., [5] proposed Naive Bayes and KNN for student placement prediction. Both are independently used to predict and compare the efficiency of Naive Bayes and KNN.

Neelam. S et. al., [6] analyzed the performance of previous-year students and predicted the placement opportunities for current-year students. This paper compares the Decision Tree and Random forest relevant to accuracy, recall, and precision. Finally, Random Forest gives better results.

Jai R and K David [7] attempted to stem a smart training data set from 4 dissimilar data sets and utilized it for the prediction process using ID3 and MLP. ID3 obtain high accuracy than MLP in all dissimilar data sets with 90% average accuracy. The obtained average accuracy of MLP is 70%. In the 4 dissimilar data sets, ID3 produced 88%, 96%, 100%, and 76% accuracy respectively and MLP produced 60%, 40%, 96%, and 84% respectively.

NT Nghe et al., [8], forecasts Undergraduate and Post Graduate students' academic level using Bayesian Network and Decision Tree. Bayesian Network works better with 3 to 12% higher accuracy than Decision Tree in different attempts like failed students and good and very good students in academics.

PROPOSED MODEL

This work proposed a voting-based ensemble model for placement prediction. This work compares the ensemble stacking and voting-based classification models with J48 for obtaining an efficient model of placement prediction. The proposed model is shown in the following Figure 1.

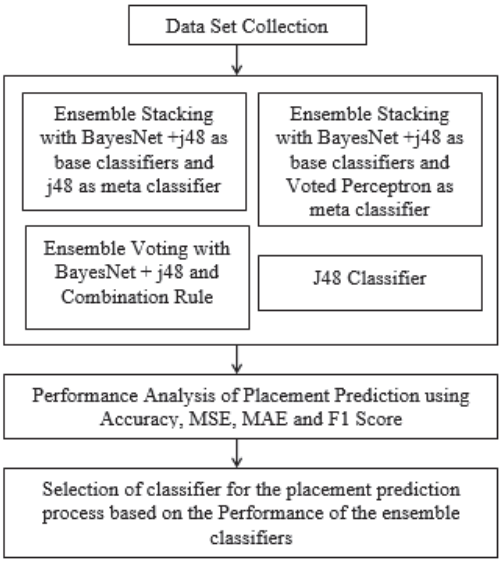

FIGURE 1. Proposed Model

Data Set collection

The data set is collected directly from the Nallamuthu Gounder Mahalingam College, placement cell, and also students. In the data set, misplaced and inappropriate data found instances are handled and define the dependent and independent attributes in the pre-processed phase.

TABLE 1. Attributes in the data set

Name	Specification	Possible Data
S No	Sequence No	Numeric Sequence
Reg No	Register No	Alphanumeric
Name	Student Name	Alphabets
Gender	Gender	{M – Male, F - Female}
Course	Name of the UG Course	{B.Sc. BCA, BCom, BA}
Year	Academic Year	{2017-2020, 2018-2021, 2019-2022}
Area of Living	Living Area of the student	{Rural, Urban}
Schooling	H.Sc. Studied School Type	{P – Private, G – Government}
Medium	Medium of study	{E-English, T-Tamil}
HSc	H.Sc. Percentage	{35% to 100%}
UG	UG Percentage	{40% to 100%}
Attendance	Attendance Percentage	{70% to 100%}
Interaction in Class	Interaction in the class hours	{Good, Average, Poor}
Paper Presented	Paper presented	{Y – Yes, N-No}
Placement Training	Placement Training Attended	{Y – Yes, N-No}
Certifications	Certifications	{Y – Yes, N-No}
CE Class Attended	Communicative English Class Attended	{Y – Yes, N-No}
Club Activities	Joined in Club Activities	{Y – Yes, N-No}
Placement	Placement Status	{Placed, Not placed}

Pre-processed data set is divided into the train (70%) and test (30%) data sets. The student placement training data set is used to fit the ML model and the test data set is used to evaluate the fitness of the ML model. Gender, course, area of living, school type and medium studied, the percentage of H.Sc., UG marks, attendance, interaction in class, placement training attended, certifications, activities of the student, and placement status is the key attributes of the student data set. The attributes handled in the student placement training and test data set are described in Table 1.

Both ensemble stacking models use BayesNet and J48 classifiers as the base classifiers. The J48 classifier was used as the meta classifier in one stacking process and the voted perceptron is used in another. In the ensemble voting model, BayesNet and J48 are used as the base classifiers and the probability average of a class of base classifiers is used for the combination rule. J48 is used to train the model individually.

Finally, above four models' classification results are compared to select the best model for placement prediction. Accuracy, MSE, MAE, and F1Score metrics are used to compare and analyze the models.

BayesNet

BayesNet is a probability-based graphical mode and is mainly utilized to calculate uncertainties using probability. A conditional probability measure is used in the BayesNet for placement prediction. Directed Acyclic graphs are used for uncertainties. BayesNet is a good one for taking an observed event and predicting the likelihood that any of the numerous known causes played a role. For example, BayesNet could reflect the probability correlations between placement status and academic data. Given a set of academic data of the students, BayesNet is used to calculate the likelihood of placement status.

J48

J48 is a tree-based classifier for machine learning and it's a very useful one to examine the category-wise and continuous data. It is a divide and conquer-based recursive strategy for determining the ideal attribute to split on at each and every stage. Information gain measure is used to select the ideal attribute at each stage. Same-class instances are treated as a leaf and labeled with the same class. IG is calculated for each attribute and used to select the best attribute to split. In the next step, entropy is calculated and finally, the best attribute will be chosen depending on the current selection parameter.

Ensemble Voting

Voting is an ensemble machine learning algorithm. In classification hard and soft voting is used. Hard voting ensemble predicts the class with the most votes by summing the votes for crisp class labels from other models. A soft voting ensemble predicts the class label with the largest sum probability by summing the predicted probabilities for class labels. In this work soft voting is used as a combination rule for an average of probability values of base classifiers J48 and BayesNet.

In this work, a voting ensemble is used to combine the predictions of J48 and BayesNet. The voting ensemble is used to enhance the performance of combined models. The average of predictions from both models is combined through the prediction of each label. Next, the majority of voted labels are selected. In the ensemble voting, both models are equally treated.

PERFORMANCE EVALUATION

In determining the best classifier and predictor model, performance evaluation metrics are needed to justify the model selection. The following metrics are used in this work to analyze and evaluate the performance of the J48, ensemble stacking, and ensemble voting models:

- Classification Accuracy
- Error Rate
- Mean Absolute Error
- Mean Squared Error
- F1 Score

Accuracy: It is a ratio of truly predicted instances over the total instances to use for prediction.

Error rate: It is a ratio of falsely predicted instances over the total instances to be used for prediction

Mean Absolute Error: It is the average value of all absolute errors. It calculates the average difference between calculated and actual values. It computes the errors between predicted values and actual values.

Mean Squared Error: It is an estimator measure based on the average squared difference between predicted values and actual values. MSE is relatively close to zero is better.

F1-Score: is a harmonic mean value of the precision and recall values.

Accuracy= (TP+TN) / (TP+ TN+FP+FN)

Recall = TP/(TP+FN)

Precision = TP/ (TP+FP)

F-Measure = (2x Precision x Recall) / (Precision + Recall)

EXPERIMENTAL RESULTS

The experimental result of the J48 classifier on training data (70% instances from student data set) is described in the following Table 2.

TABLE 2. Performance of j48

Accuracy	MSE	MAE	TP	FP	F1
87.5	0.3002	0.1279	0.875	0.263	0.884

The J48 classifier produced 87.5% of accuracy with a 12.5% of error rate and 0.884 as F1Score. The MSE rate is 0.3002 and the MAE rate is 0.128.

The experimental results of the ensemble stacking model with BayesNet and 48 as base classifiers and J48 as meta classifiers used for the student training data are described in the following Table 3.

TABLE 3. Performance of Ensemble Stacking1

Accuracy	MSE	MAE	TP	FP	F1
89.29	0.2966	0.1029	0.893	0.260	0.898

This ensemble stacking model produced 89.29% of accuracy with a 10.71% of error rate and a 0.898 F1 Score. The MSE rate is 0.2969 and the MAE rate is 0.1029.

The experimental results of the ensemble stacking model with BayesNet and 48 as base classifiers and Voted Perceptron as meta classifiers used for the student training data are described in the following Table 4.

TABLE 4. Performance of Ensemble Stacking2

Accuracy	MSE	MAE	TP	FP	F1
83.93	0.3958	0.1587	0.839	0.513	0.844

This ensemble stacking model produced 83.93% of accuracy with a 16.07% of error rate and 0.844 F1 Score. The MSE rate is 0.3958 and the MAE rate is 0.1587

The experimental results of the ensemble voting model with BayesNet and 48 as base classifiers and the probability average for the combination rule used for the student training data are described in the following Table 5.

TABLE 5. Performance of Ensemble Voting model

Accuracy	MSE	MAE	TP	FP	F1
91.07	0.298	0.1337	0.911	0.013	0.920

This ensemble voting model produced 91.07% of accuracy with an 8.93% of error rate and a 0.920 F1 Score. The MSE rate is 0.2988 and the MAE rate is 0.1337.

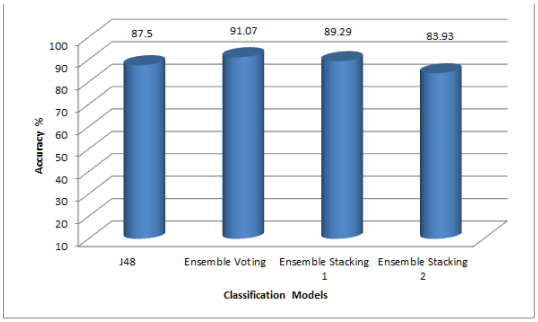

FIGURE.2. Classification Accuracy

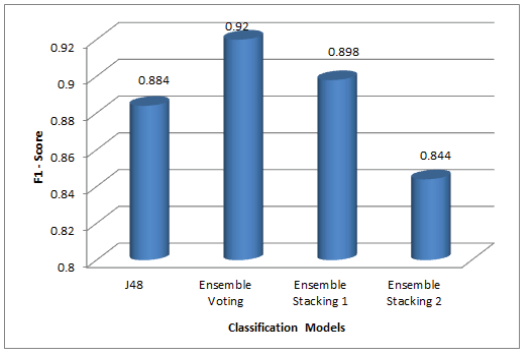

FIGURE 3. F1 –Score of classifiers

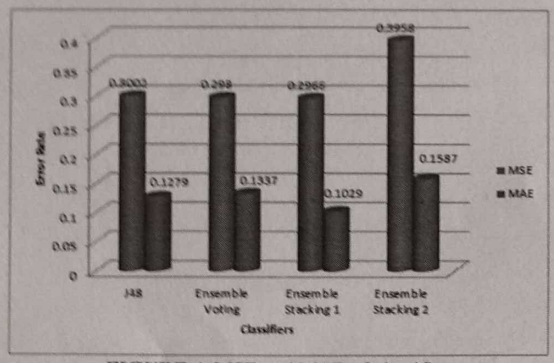

FIGURE 4. MSE and MAE of classifiers

The above figures (Figures 2 to 4) indicate that the proposed ensemble voting-based classifier (BayesNet + J48) with Average of probabilities as the combination rule for both classifiers outperforms the other ensemble stacking classifier models (BayesNet + J48 as the base and J48 / Voted Perceptron as the meta classifier).

CONCLUSION

Ensemble voting with (BasyesNet+J48) is the most suitable model for student placement prediction. This model produces high classification accuracy (91.07%) and F1-Score (0.920) than different ensemble stacking models and the j48 classifiers. The proposed ensemble voting (BayesNet + j48) produces a lower error rate, lower mean absolute error (0.1337), and lower mean squared error rate (0.298) than the other two models. The proposed ensemble voting (BayesNet + j48) helps to use efficient measures for the suitable placement of students. However, integrating other constraints such as failures in the semester examinations, and student activity/behavior in the campus-related attributes are included in the classification to increase the performance of the proposed model.

REFERENCES

1. S. Dutta and S. K. Bandyopadhyay, Asian Journal of Research in Computer Science **5**, 1-12 (2020).
2. B. Sen, E. Ucar and D. Delen, Expert Sys with Applications **39**, 9468-9476 (2012).
3. D. Thammasiri, N. Ksasp, P. Meesed and D. Delen, Expert Systems with Applications **41**, 321-330 (2014).
4. VK Harihar and DG Bhalke, A Journal of Physical Sciences, Engineering and Technology **12**, 85-91 (2020).
5. H Shreyas, P Aksha and H Suma, International Research Journal of Engineering and Technology **6**, 4577-4579 (2019).
6. Neelam S, Intl. J. of Applied Eng. Res. **14**, 2188-2191 (2019).
7. R Jai and K David, International Journal of Trend in R & D **17**, 33-36 (2021).
8. N Thai Nghe, P Janecek and P Haddawy, Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference, 7-12 (2007).