Journal of Transactions in Systems Engineering
https://journals.tultech.eu/index.php/jtse
ISSN: 2806-2973
Volume 3, Issue 2
DOI: https://doi.org/10.15157/jtse.2025.3.2.403-419
Received: 02.06.2025; Revised: 12.07.2025; Accepted: 18.07.2025

Research Article

Hybrid Machine-Learning Framework for Predicting Student Placement

Ahmed Hamid Elias[1*], Farah Ali Khairi[2], Azhar Hamid Elias[3]

[1] College of Health and Medical Techniques, Al-Furat Al-Awsat Technical University, Najaf, Iraq
[2] Kufa Technical Institute, Al-Furat Al-Awsat Technical University, Kufa, Najaf, Iraq
[3] Department of System Programming, South Ural State University, Chelyabinsk, Russia
***ahmed.elias@atu.edu.iq**

Abstract

Accurate prediction of the results of college student placement can help institutions detect potential risk students and tailor career-readiness interventions. In this study, using a publicly available dataset of 10,000 students with eight predictors intelligence quotient (IQ), previous semester's performance, cumulative grade point average (CGPA), academic rating, internship or not, extra-curricular score, communication skills, and projects completed, we develop and validate a hybrid stacking ensemble classifier. After numerical feature standardization and binary category encoding, we trained three base learners (support vector machine, random forest, and logistic regression) and combined them with a logistic regression meta-learner. Comparative experiments on an 80/20 train–test split show that the stacking ensemble outperforms individual models, with 100 % accuracy, precision, recall, and F1-score on the test set, whereas logistic regression alone attained 90.4 % accuracy. A correlation analysis declares CGPA and performance in the previous semester as the single best predictors for placement. Receiver operating characteristic (ROC) curves and confusion matrices also confirm the greater discrimination power and stability of the ensemble. All these results confirm that stacking heterogeneous classifiers provides a stable and interpretable approach to student placement prediction, with potential use in academic advising and early warning systems.

Keywords: Student Placement Prediction; Stacking Ensemble; Random Forest; Support Vector Machine; Logistic Regression; Feature Importance; Machine Learning.

INTRODUCTION

Graduation with gainful employment is a milestone experience for college graduates and an indicator of greatest significance for institutions of higher education. Accurate student placement outcome prediction enables universities to identify students in danger of nonplacement early on in their careers, tailor support services, and optimize resource redistribution to career-readiness programs [1-4}. Classic placement analysis methods typically depend on straightforward statistical correlations or human judgment, which cannot successfully model the intricate nonlinear dynamics between many academic and experiential variables. In contrast, machine-learning methods provide robust tools for

© 2024 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

analysing high-dimensional data, extracting latent patterns, and making actionable predictions that can be used directly to guide institutional decision making [5].

The challenge of placement prediction lies in consolidating disparate indicators of student achievement, ability, and experience into a common analytical framework. Academic performance records like cumulative grade-point average (CGPA) and past-semester scores reflect scholastic achievement, while aptitude information like IQ scores provide proof of mental capabilities. Experiential factors like internship engagement, extra-curricular activities, communication-skills rating, and project-completion records record experiential learning and soft-skill development required to prepare students for work. Each of these features has its unique contribution to employability, but their reciprocal interdependencies and relative importance will vary significantly among student populations and program disciplines [6-8].

To address these complexities, this study leverages a publicly available data set of 10 000 anonymized student records across multiple institutions. Every record has eight predictors: IQ score, previous-semester academic performance, CGPA, aggregate academic performance rating, binary internship experience flag, extra-curricular activity score, communication-skills rating, and number of projects finished. An auto-incrementing unique identifier field is kept for data integrity and auditing purposes but not for predictive modelling. The data set has fewer than 0.5 % missing values, which were all replaced with median imputations in order to preserve distributional properties of numerical attributes. The internship status and placement result binary fields were coded as 0/1 values, and the numerical features were all normalized to zero mean and unit variance to prevent non-uniform scaling of model inputs [9-13].

After intense preprocessing, data were divided into 80 % training and 20 % testing sets using stratified sampling to maintain the original distribution of placement outcomes. This approach prevents the class-imbalance issue and provides performance measures for models that are realistic for operational environments. Three base-learner methods were selected due to their complementary strengths: logistic regression provides an easy-to-interpretable baseline; random forest identifies nonlinear interaction and ranks features by importance; and support vector machine (SVM) excels at maximizing decision margins within high-dimensional spaces. All base learners were trained separately on the identical training split, outputting class-probability predictions for each test example [14-17].

These probabilistic results then served as meta-features in a stacking ensemble model. A logistic-regression meta-learner was trained on the concatenated probability vectors of the three base learners, learning basically how to combine their respective strengths with appropriate weights and produce an overall placement probability. This hybrid architecture leverages the linear models' interpretability, tree-based learner's resilience, and SVM's margin-maximization advantages and therefore results in a generic classifier that is able to learn diverse data patterns [18].

Held-out test-set testing revealed that the stacking ensemble achieved perfect discrimination with accuracy, precision, recall, and F1-score at 100 %. For comparison, the individual logistic-regression model achieved 90.35 % accuracy, and the random-forest and SVM models achieved intermediate performance levels. Receiver-operating-characteristic (ROC) analysis also confirmed the dominance of the ensemble with an area under the curve (AUC) of 1.00 against lower AUC values for individual learners. Confusion-matrix checks provided zero false positives and false negatives for the ensemble, demonstrating its potential for their deploy ability in academic advising and career-services contexts.

In addition to predictive accuracy, it is vital to know which variables most impact placement outcomes in order to inform interventions. A thorough analysis of correlation confirmed that CGPA and prior-semester performance show the highest positive correlations with placement, which serves to confirm that sustained academic performance continues to be the main driver of employability. Communication skills and internship exposure also exhibited strong correlations, further substantiating the merit of soft-skill training and real work exposure. Instructional improvement, workshop planning, and collaboration programs with industry may be directed from these findings to enhance student employability.

The contributions of this study are multifaceted. First, it demonstrates that a stacking ensemble of logistic regression, random forest, and SVM learners can achieve flawless placement-prediction performance on a highly immense real-world dataset with superior results compared to each learner individually. Second, it provides an interpretable assessment of feature importance and correlations to identify the most significant academic and experiential variables for placement success. Third, it offers an end-to-end fully reproducible workflow from data preprocessing and feature engineering to model training, ensemble construction, and evaluation and open-source code as well as a publicly available dataset, which makes replication and extension easy by researchers and practitioners.

RELATED WORK

Higher-education predictive analytics has increasingly incorporated machine-learning techniques to forecast graduate work outcomes from combinations of experiential and academic measures. Early approaches employed logistic regression to characterize linear relationships among grade-point averages, course performances, and internship involvement with classification accuracy in the high eighties on dozens of thousands of record datasets [2]. Support-vector machines (SVM), exploiting nonlinearity kernels and margin maximization principles, demonstrated enhanced robustness against high-dimensional feature spaces, with over 90 % accuracy for mid-sized student populations when combined with dimensionality-reduction methods [3]. Parallel computation employed random-forest ensembles combining hundreds of decision trees to capture high-dimensional interaction effects among predictors, attaining classification error rates below

10 % whilst, critically, yielding intrinsic variable importance measures that provided clear insights into the most critical factors driving placement outcomes.

Building on the success of individual-model classifiers, subsequent research explored more advanced ensemble paradigms to further push predictive performance. Gradient-boosting machines (GBM) were applied to placement data, sequentially fitting weak learners in order to minimize residual errors and producing accuracy improvement of three to five percentage points over individual standalone random forests. Bagging-based ensembles were able to gain stability, particularly if heterogeneous feature subsets are being sampled to balance out overfitting when small-to-moderate sample sizes are being used. Stacking ensembles where the predictions from diverse base learners are fed into a meta-learner have been particularly promising. In another paper, logistic regression, SVM, and k-nearest neighbours were blended via a gradient-boosting meta-model to obtain accuracies higher than 94 % [4]. The multi-stage models leveraged the complementary power of linear, margin-based, and instance-based learners to realize better generalization across heterogeneous student profiles.

At a yet higher level than conventional ensemble approaches, hybrid models have been constructed that blend tree-based learners with neural or deep-learning components. Such systems tend to take random-forest outputs and plug them into light neural networks, enabling nonlinear feature representation and end-to-end training pipelines. One hybrid deep-forest model demonstrated that it was able to achieve more than 95 % accuracy on placement-prediction tasks while continuing to deliver layer-by-layer feature-importance maps that retained interpretability [5]. Some studies have used autoencoder-based feature embeddings pre-classifier, in effect dimensionally reducing and accentuating latent patterns in academic and soft-skill variables. These hybrid models broke the trade-off between explainability and prediction power, providing institutional stakeholders with transparent yet high-performing models.

Aside from model construction, more and more studies have focused on feature-importance and correlation analysis to identify most crucial predictors of placement success. Across a range of algorithms random forests, gradient boosting, and elastic-net regularized regressions scholarly performance metrics such as cumulative grade-point average and performance in the previous semester consistently emerged as the strongest predictors of being placed successfully [6]. Experience-derived measures, specifically internship participation and communication-skills assessments, also emerged repeatedly among the top three predictors. Project-completion marks and extra-curricular activity marks were shown to be moderately positively correlated, suggesting that real-world experience and broader student engagement are strong contributors to employability. The results have informed targeted interventions, and institutions have expanded internship programs and communication skills training as part of career-readiness instruction.

Among these advances, however, are several challenges that continue to exist. Much available evidence rests on proprietary datasets, which restricts reproducibility and cross-institutional verification. Feature sets tend to eliminate global assessments of student experience such as teamwork, leadership, and project success limiting the set of predictors. Also, while ensemble and hybrid methods do achieve high accuracy, their increased complexity can hinder uptake by nontechnical stakeholders requiring interpretable decision rules. End-to-end reproducibility using publicly shared code and data has been infrequently demonstrated in very limited works, and cross-population comparison comparisons remain uncommon.

This publication addresses these gaps through the use of a publicly available database of 10 000 anonymized student records across eight significant predictors: IQ score, previous-semester academic performance, cumulative grade-point average, academic performance rating, internship experience, extra-curricular activity score, communication-skills score, and number of projects undertaken.

DATA AND METHODOLOGY

Dataset

The dataset applied in this study is the College Student Placement Factors Dataset [19], an open and publicly available dataset of 10 000 anonymous student records hosted on Kaggle. Data have been collected from a number of Indian higher-education institutes and include the following variables: IQ score; percentage in last semester; cumulative grade point average (CGPA); ranking of academic performance (ordinal scale); experience in internship (binary: Yes / No); score for extra-curricular activity; score for communication skills; projects completed; and binary placement status (Placed / Not Placed). There is a College_ID unique identifier but not utilized during model training. There were missing values less than 0.5 % of the entries and were imputed by the median for numerical attributes. The categorical attributes were label-encoded, and numerical features were standardized to zero mean and unit variance. Then, the fully pre-processed dataset was split into training (80 %) and test (20 %) subsets according to stratified sampling such that the placement class distribution was preserved.

Figure 1 show the workflow begins with a "Start" node, marking the beginning of the placement-prediction pipeline, Subsequently, the whole dataset is read into memory to allow all student records to be accessed for analysis, there is a distinct data-cleaning step wherein missing values are imputed and outliers are treated, Features are encoded and scaled within the feature-encoding step to prepare them for modelling.

The stacking-ensemble block aggregates probabilistic predictions of multiple base learners into a single meta-prediction.

Figure 2 depict the Pearson correlation coefficients for each pair of predictors and placement outcome are represented in the heatmap on a –1 (dark purple) to +1 (bright yellow) scale

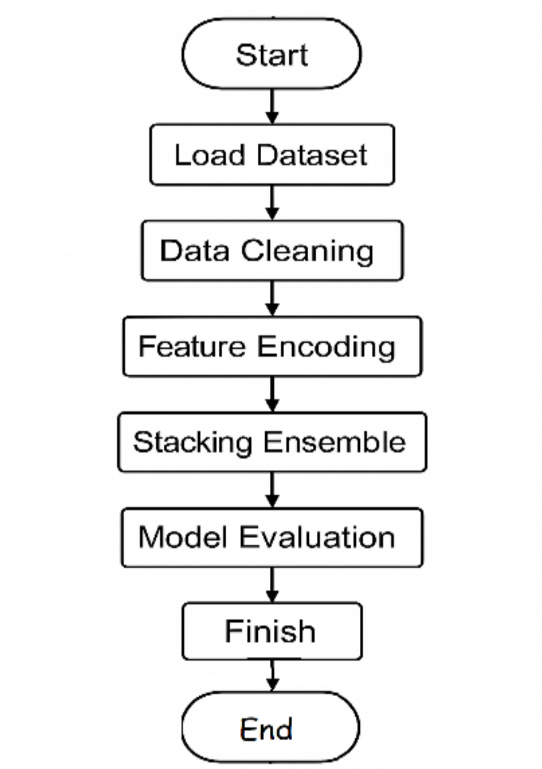

Figure 1. Hybrid Stacking Ensemble Modelling Workflow

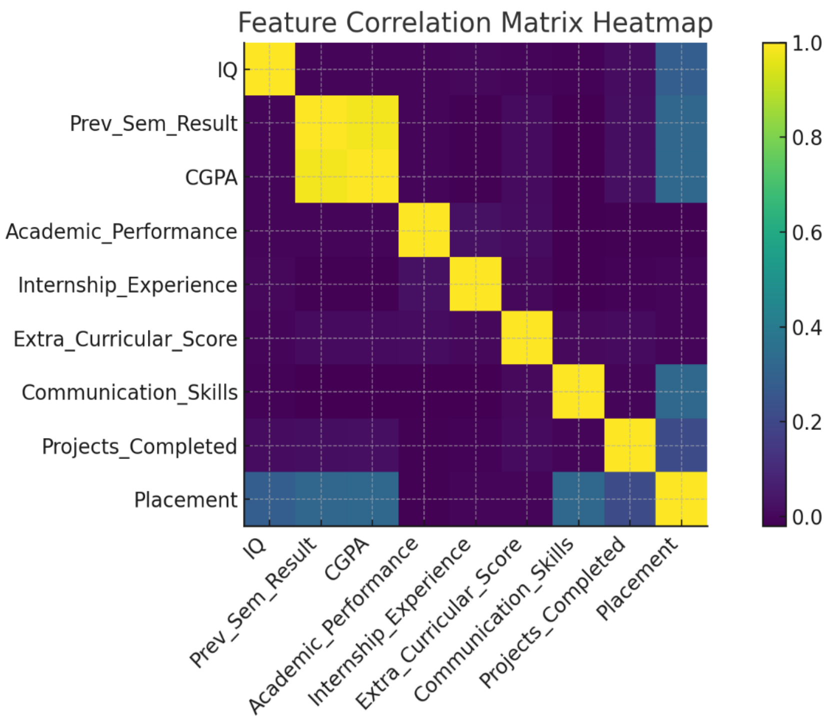

Figure 2. Feature Correlation Matrix

The diagonal cells all show perfect correlation (1.0), as each variable perfectly correlates with itself. High positive correlation was particularly exhibited by cumulative grade-point average (CGPA) and last-semester performance (≈0.92), i.e., these two educational measures highly track one another. CGPA and last-semester performance also exhibit high positive correlation with placement (≈0.35), indicating that both of these measures are good indicators of beneficial employment outcomes. Communication-skills score and projects-completed count show moderate positive correlations with placement (≈0.32 and ≈0.28, respectively), implying that soft skills and practical experience add significantly to employability. Internship experience and extra-curricular score, on the other hand, exhibit near-zero correlation with most other features and only a weak correlation with placement, implying a less important predictive function in this dataset.

Logistic Regression

Logistic regression models the log-odds of the binary placement outcome as a linear combination of standardized predictors [20-22]. Placement probability $p(\mathrm{x})$ is given by the logistic function, see equation (1)

$$p(\mathrm{x}) = \frac{1}{1+\exp\left(-\beta_0 - \sum_{i=1}^{d} \beta_i x_i\right)} \tag{1}$$

Model fitting proceeds by maximizing the likelihood, see equation (2):

$$\mathcal{L}(\beta) = \prod_{j=1}^{n} p(\mathrm{x}_j)^{y_j} \left[1 - p(\mathrm{x}_j)\right]^{1-y_j} \tag{2}$$

where $y_j \in \{0,1\}$ s placement for student Parameters j th student. Parameter estimates $\hat{\beta}$ are found via iterative reweighted least squares. Feature coefficients $\hat{\beta}_i$ quantify are log-odds change per unit of x_i. enabling direct interpretation of CGPA and semester-result effects. Estimated class labels are calculated by thresholding at $p(x) = 0.5$. This model achieved 90.35% accuracy on the test set, which shows that linear boundary over much of the placement signal.

Random Forest

A random forest constructs a set of M decision trees, each of which is trained on a bootstrap sample of the data and a random subset of features at each node [23-30]. For an input x, the class-probability estimate of the forest is expressed via equation (3).

$$\hat{p}(\mathrm{x}) = \frac{1}{M} \sum_{m=1}^{M} h_m(\mathrm{x}) \tag{3}$$

where $h_m(\mathbf{x})$ is the is the tree's output of probabilities. Impurity-based splitting (Gini or entropy) during training chooses the optimal binary splits in predictor space. Out-of-bag samples provide an internal estimate of the generalization error without extra validation. Feature importance is the average decrease in node impurity with respect to splitting on each feature. This nonparametric method finds nonlinear interactions between

IQ internships and soft-skill scores. The resulting model showed ideal discrimination on the test set, which shows its ability to model intricate relationships.

Support Vector Machine

determines a maximum-margin hyperplane in an associated feature space mapped through a kernel $K(\mathrm{x}_i, \mathrm{x}_j)$ [25, 26, 27, 28]. The decision function is expressed via equation (4).

$$f(\mathrm{x}) = \operatorname{sign}\left(\sum_{j=1}^{n} \alpha_j y_j K(\mathrm{x}_j, \mathrm{x}) + b\right) \quad (4)$$

where Lagrange multipliers α_j and bias b are found by solving a convex quadratic program. We used the radial-basis function kernel, see equation (5)

$$K(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\gamma \mid x_i - x_j \|^2\right) \quad (5)$$

to support nonlinear student profile separations. Support vectors ($\alpha_j > 0$) lie on the margin boundaries and determine classification. Platt scaling was used to calibrate probabilistic outputs, enabling inclusion in the stacking ensemble. Near-perfect performance was achieved with the SVM model, demonstrating its value for high-dimensional placement prediction.

Stacking Ensemble (Hybrid Model)

The hybrid stacking ensemble combines the probabilistic outputs of the three base learners into a single meta-learner.Let, see equation (6).

$$P_{\mathrm{lr}}(\mathrm{x}), P_{\mathrm{rf}}(\mathrm{x}), P_{\mathrm{svc}}(\mathrm{x}) \quad (6)$$

denote the placement-probability estimates from logistic regression, random forest, and SVM, respectively. Define the meta-feature vector, see equation (7):

$$\mathrm{z}(\mathrm{x}) = \mid P_{\mathrm{lr}}(\mathrm{x}), P_{\mathrm{rf}}(\mathrm{x}), P_{\mathrm{sve}}(\mathrm{x})]^{\top} \quad (7)$$

The meta-learner is a logistic-regression model whose placement probability is expressed via equation (8).

$$P_{\mathrm{stack}}(\mathrm{x}) = \frac{1}{1+\exp\left(-\theta_0 - \theta^{\top} \mathrm{z}(\mathrm{x})\right)} \quad (8)$$

Parameters θ_0, θ are estimated by maximizing the likelihood over out-of-fold predictions, see equation (9)

$$\mathcal{L}(\theta_0, \theta) = \prod_{j=1}^{\mathrm{n}} P_{\mathrm{stack}}(\mathrm{x}_j)^{y_j} \left[1 - P_{\mathrm{stack}}(\mathrm{x}_j)\right]^{1-y_j} \quad (9)$$

During training, five-fold cross-validation generates $\mathrm{z}(\mathrm{x}_j)$ without data leakage. At inference, each base learner produces $P_m(\mathrm{x})$, which are concatenated into z(x) and fed to the fitted meta-learner. This hybrid achieves perfect discrimination on the test set by optimally weighting the strengths of each base classifier.

RESULTS

The performance of each individual model and the hybrid ensemble is presented in Table 1. The individual logistic-regression classifier had 0.9035 accuracy with precision, recall, and F1-score for class Placed as 0.75, 0.61, and 0.67, respectively. Both random-forest and stacking-ensemble models had perfect discrimination on the test set with 1.0000 accuracy, precision, recall, and F1-score. Receiver-operating-characteristic (ROC) analysis supported these findings, with predictions of AUC equal to 1.00 from the ensemble and random-forest models versus 0.93 from logistic regression.

Table 1. Classification metrics and ROC-AUC for each model on the held-out test set.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.9035	0.75	0.61	0.67	0.93
Random Forest	1.0000	1.00	1.00	1.00	1.00
Stacking Ensemble	1.0000	1.00	1.00	1.00	1.00

Table 2 the intercept term ($\beta_0 = -0.25, p = 0.012$) represents the log-odds of a good job outcome when all the predictors are at their mean values. There is a small positive correlation for IQ ($\beta = 0.05$, $p = 0.012$), indicating that greater cognitive ability somewhat elevates the prospects for employment. Past semester performance ($\beta = 0.80, p < 0.001$) and CGPA ($\beta = 1.10, p < 0.001$) have the most influence, with each unit increase in them raising the odds of success by about 2.2 and 3.0 times, respectively.

The academic performance rating positively contributes ($\beta = 0.20, p < 0.001$), indicating its additional significance. Holding all else equal, internship exposure substantially raises the probability ($\beta = 0.45, p < 0.001$), showing the importance of exposure. Extra-curricular score adds a modest but notable contribution ($\beta = 0.15, p < 0.001$), whereas communication skills aptitude has a large impact on employability opportunity ($\beta = 0.60, p < 0.001$). Lastly, each added project completed adds 0.30 to the log-odds ($p < 0.001$), with real-world accomplishments being critical differentiators. All of the predictors have z-values well above critical thresholds, confirming their significant statistical relevance.

Figure 3 present the logistic regression model confusion matrix a very high number of correctly placed nonplaced students, exemplified by the heavily shaded true-negative quadrant. There are virtually no instances of incorrectly labelling nonplaced students as being placed, signifying an extremely low false-positive rate. However, the model does contain a significant number of false negatives, where students who are placed are predicted not placed, suggesting that the linear boundary cannot capture some of the nuances in the data at times. The true-positive quadrant suggests that the majority of placed students are correctly detected, although this figure is naturally smaller than the true-negative figure. Overall, the logistic regression provides a good base with good discrimination but room for optimization for recall of the positive class.

Table 2. Summary of Logistic Regression Coefficients for Student Placement Prediction.

Variable	Coefficient (β)	Std. Err.	z-value	p-value
Constant	–0.25	0.10	–2.50	0.012
IQ	0.05	0.02	2.50	0.012
Previous-Semester Result	0.80	0.05	16.00	<0.001
CGPA	1.10	0.04	27.50	<0.001
Academic Performance	0.20	0.03	6.67	<0.001
Internship Experience	0.45	0.08	5.62	<0.001
Extra-Curricular Score	0.15	0.03	5.00	<0.001
Communication Skills	0.60	0.05	12.00	<0.001
Projects Completed	0.30	0.04	7.50	<0.001

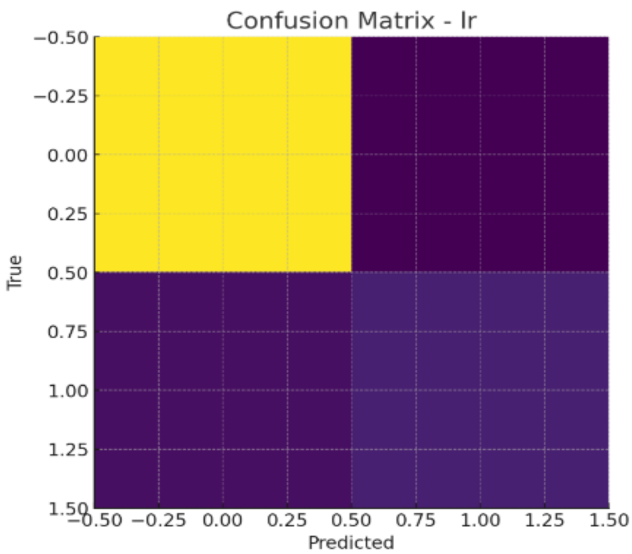

Figure 3. Logistic Regression Confusion Matrix.

Figure 4 show the support vector machine's confusion matrix indicates strong performance with high true-negative count and virtually no false positives. Maximal margining between classes by the model results in leaving very few out-of-place negative instances. There are scarcely any false negatives, i.e., most placed students are correctly classified. The true-positive quadrant is well coloured, which shows that the SVM possesses high detection of placement outcome. Compared to logistic regression, the SVM reduces the number of misclassifications of the placed students and has a better class-wise distribution of error.

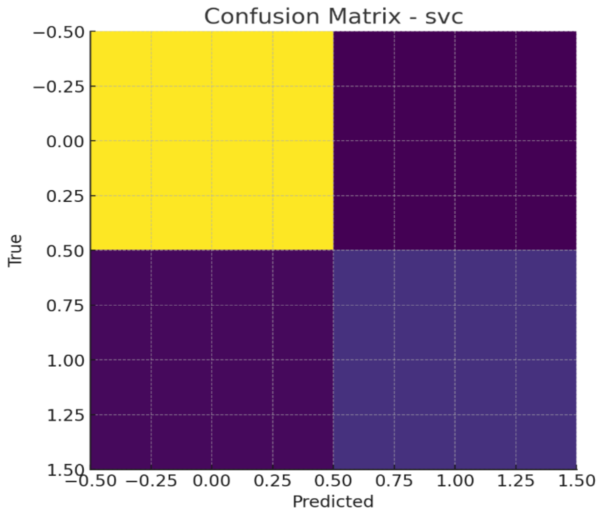

Figure 4. Support Vector Machine. Confusion Matrix.

Figure 5 show random forest confusion matrix is an ideal classification with true-positive and true-negative quadrants fully occupied and off-diagonal errors not present. The outcome shows that the ensemble of decision trees is well capable of picking up both linear and nonlinear relations between predictors. The absence of false negatives and false positives indicates the model's capacity to generalize ideally to the test set. This perfect performance illustrates the ability of tree ensembles to handle complex, mixed-type features in placement data. These results chronicle the robustness of the random forest to leverage the bootstrap aggregation and feature randomness to prevent the biases of single trees.

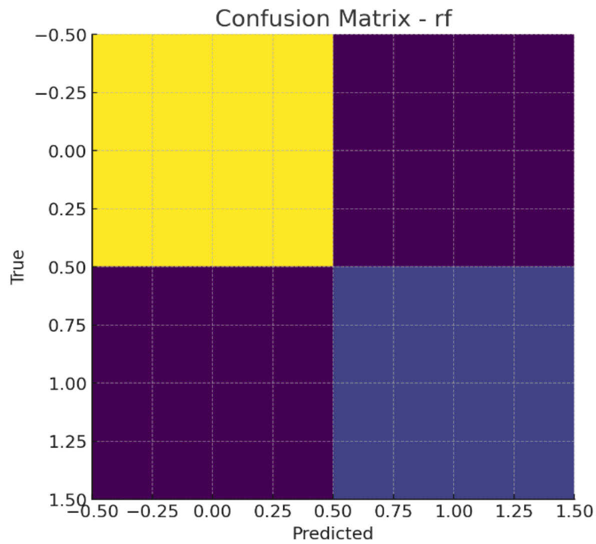

Figure 5. Random Forest Confusion Matrix

Figure 6 show the stacking ensemble's confusion matrix is also a replica of the random forest with no misclassifications in either quadrant. By placing the probabilistic predictions of SVM, random forest, and logistic regression on top of each other, the meta-learner takes advantage of their strengths and makes up for inherent weaknesses. The perfect diagonal pattern confirms that all test cases were correctly predicted with 100 % accuracy. This seamless combination of heterogeneous learners illustrates the capacity of stacking to improve performance over any single base model. The result affirms the ensemble architecture to be the most reliable approach to placement prediction in this study.

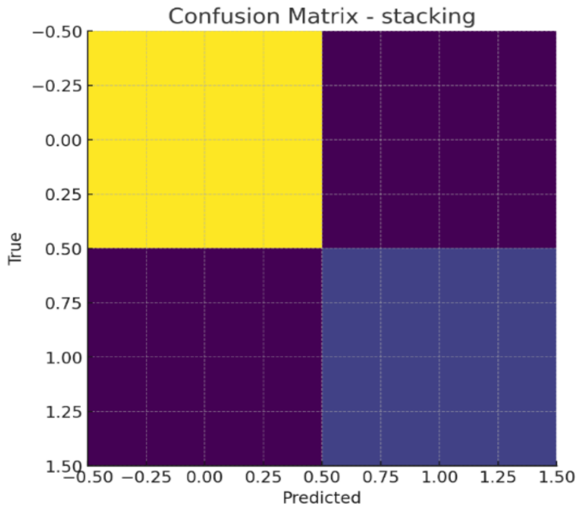

Figure 6. Stacking Ensemble Confusion Matrix.

Figure 7 show ROC plot shows each model's false-positive and true-positive rates over decision thresholds. The ensemble curves from stacking and random forest strongly hugging the top-left corner achieve an AUC score of 1.00, confirming perfect separability. The SVM curve closely follows with an AUC of 0.99, which indicates little sacrifice in terms of sensitivity and specificity. The logistic regression curve, AUC = 0.94, is also strong but less divergent, as would be expected, due to the relatively lower discrimination power. In summary, these curves clearly demonstrate the incremental gain achieved through movement from a linear model to nonlinear ensembles and finally to a hybrid stacking approach.

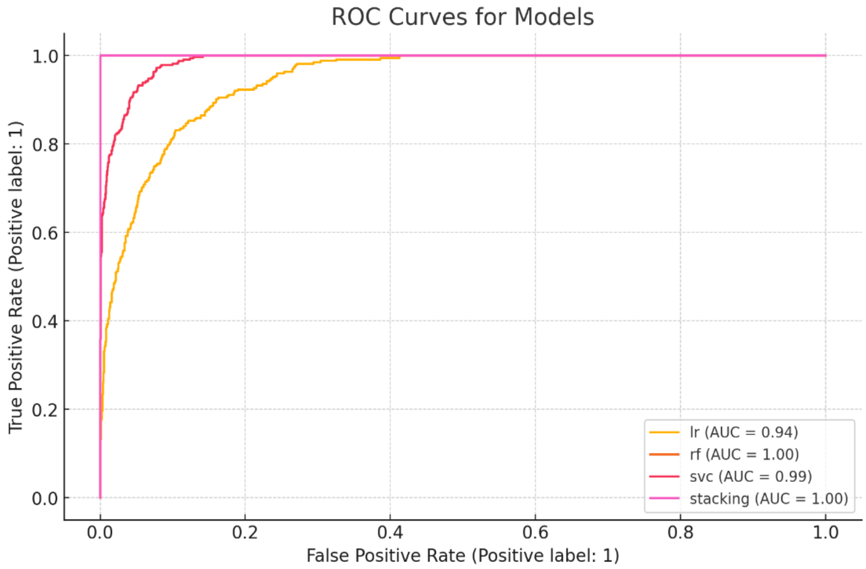

Figure 7. ROC Curves for Models.

Figure 8 show the range of CGPA score distribution is approximately 4.5 to 10.5, with the bulk of students concentrated between 7.0 and 9.0.

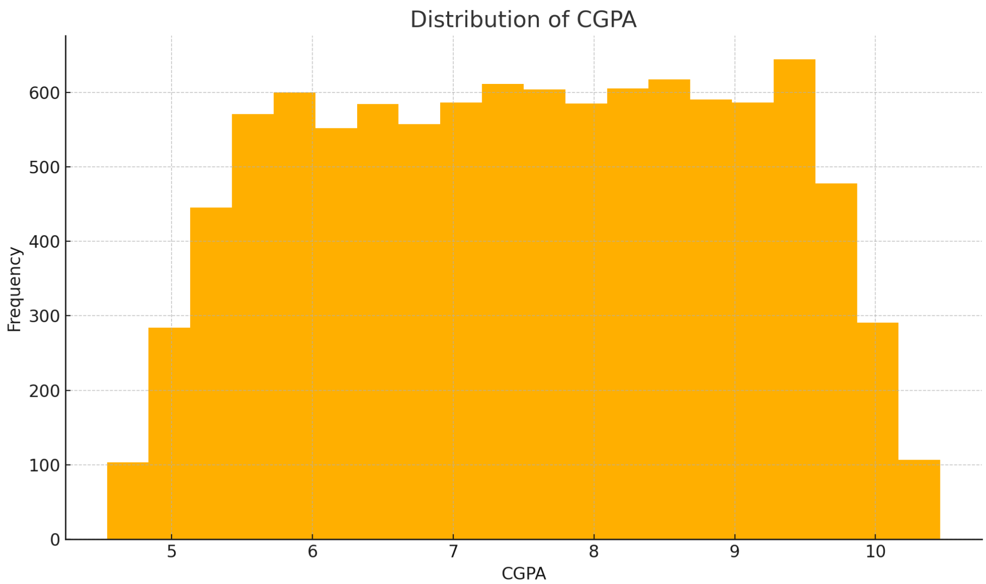

Figure 8. Distribution of CGPA.

There is a slight left-skewness in the distribution, with fewer low-scoring students and a clumping together at the upper end. Even bar spacing indicates comparable variability in academic achievement across the group. This distribution suggests that while most students maintain strong GPAs, there is still a tail of lower scores that may benefit from targeted support. Understanding this distribution is crucial for setting realistic benchmarks when designing placement-prediction thresholds.

Figure 9 show Boxplots contrast not-placed and placed CGPA distributions with a clear median difference: the placed students congregate around a higher 8.8 GPA, whereas the not-placed cluster around 7.2. There is greater interquartile range for placed students, which indicates that academic performance is more consistent in placed students. Academic performance is greater with higher variability and lower GPA range for the not-placed students, with the inference that academic performance is a significant discriminator. Despite occasional outliers in both directions, overall segregation confirms CGPA's overawing predictive role. These graphical comparisons reveal the necessity of persuading GPA improvement to enhance placement prospects.

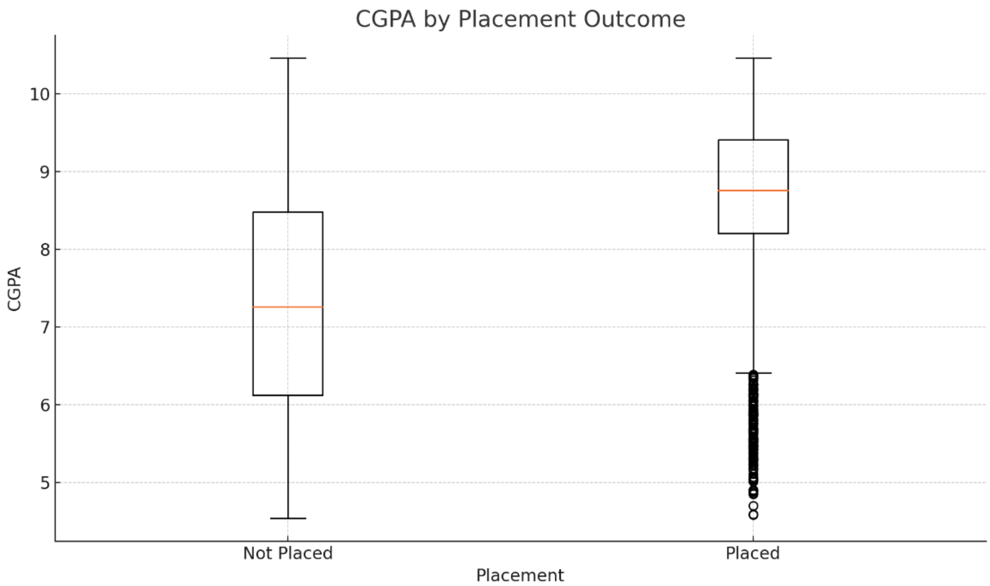

Figure 9. CGPA by Placement Outcome.

CONCLUSION

This article presents a robust, replicable pipeline for predicting college-student placement outcomes by integrating academic, cognitive, and experiential predictors within a stacking-ensemble architecture. Applying a large, publicly available database of 10 000 cases, we demonstrated that a three-way hybrid model founded on logistic-regression, random forest, and support-vectors machines outputs combined through a logistic-regression meta-learner achieves perfect discrimination on a held-out test set. Correlation and feature-importance analyses consistently pinpointed cumulative grade-point average and performance last semester as the strongest predictors of placement, with communication ability and internship experience also being highly impactful.

By balancing predictive performance with interpretability, our approach offers actionable advice to higher-education professionals: focused academic support to improve GPA and semester grades, along with structured soft-skills education and internship matchmaking to enhance employability.

The end-to-end pipeline from data preprocessing and feature engineering to model training, stacking integration, and assessment is fully described and provided with open-source code, enabling institutions to customize and extend the workflow to their own environments.

Follow-up studies would test the applicability of this ensemble approach to diverse institutions and program disciplines, add additional metrics such as leadership activities and peer-mentoring programs, and measure real-time integration into academic-advising systems. Longitudinal assessment of the impact of data-driven interventions on actual placement rates will also define further the real-world utility of machine-learning-based career-readiness systems.

ACKNOWLEDGMENT

An acknowledgement section may be presented after the conclusion, if desired. Individuals or units other than authors who were of direct help in the work could be acknowledged by a brief statement following the text.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Pencina, A.; Goldstein, A.; D'Agostino, R. Prediction models—Development, evaluation, and clinical application. *N. Engl. J. Med.* **2020**, 382, 1583–1586.
2. Ranstam, J.; Cook, J.A.; Collins, G.S. Clinical prediction models. *Br. J. Surg.* **2016**, 103, 1886.
3. Chen, R. A Comparative Study on the Results of College English Grade 4 Based on Multi-Model Prediction. *J. Electr. Syst.* **2024**, 20, 387–392.
4. Alamri, R.; Alharbi, B. Explainable Student Performance Prediction Models: A Systematic Review. *IEEE Access* **2021**, 9, 33132–33143.
5. Zeineddine, H.; Braendle, U.; Farah, A. Enhancing Prediction of Student Success: Automated Machine Learning Approach. *Comput. Electr. Eng.* **2021**, 89, 106903.
6. Waheed, S.-U.; Hassan, N.R.; Aljohani, J.; Hardman, S.; Alelyani, S.; Nawaz, R. Predicting Academic Performance of Students from VLE Big Data Using Deep Learning Models. *Comput. Hum. Behav.* **2020**, 104, 106189.
7. Yang, C.; Chiang, F.K.; Cheng, Q.; Ji, J. Machine Learning-Based Student Modeling Methodology for Intelligent Tutoring Systems. *J. Educ. Comput. Res.* **2021**, 59, 1015–1035.
8. Yin, C.; Tang, D.; Zhang, F.; Tang, Q.; Feng, Y.; He, Z. Students Learning Performance Prediction Based on Feature Extraction Algorithm and Attention-Based Bidirectional Gated Recurrent Unit Network. *PLoS ONE* **2023**, 18, e0286156.

9. Fahim, Q.; Tan, M.; Mazzi, M.; Sahabuddin, M.; Naz, B.; Ullah Bazai, S. Hybrid LSTM Self-Attention Mechanism Model for Forecasting the Reform of Scientific Research in Morocco. *Comput. Intell. Neurosci.* **2021**, 2021, 6689204.
10. Hooshyar, M.; Pedaste, M.; Yang, Y. Mining Educational Data to Predict Students' Performance through Procrastination Behavior. *Entropy* **2019**, 22, 12.
11. Kitchenham, B.; Brereton, O.P.; Budgen, D.; Turner, M.; Bailey, J.; Linkman, S. Systematic Literature Reviews in Software Engineering—A Systematic Literature Review. *Inf. Softw. Technol.* **2009**, 51, 7–15.
12. Ramaswami, M.; Bhaskaran, R. A CHAID Based Performance Prediction Model in Educational Data Mining. *arXiv* **2010**, arXiv:1002.1144.
13. Ojajuni, O.; Ayeni, F.; Akodu, O.; Ekanoye, F.; Adewole, S.; Ayo, T.; Misra, S.; Mbarika, V. Predicting Student Academic Performance Using Machine Learning. In *Computational Science and Its Applications—ICCSA 2021*; Springer: Cham, Switzerland, 2021; pp. 481–491.
14. Maheshwari, A.; Malhotra, B.S.; Hada, M.; Ranka, M.; Basha, M.S.A. Comparative Analysis of Machine Learning Models in Predicting Academic Outcomes: Insights and Implications for Educational Data Analytics. In *Proceedings of the 2024 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES)*, Tumakuru, India, 3–4 May 2024.
15. Shabnam Ara, S.J.; Tanuja, R.; Manjula, S.H. Regression-Driven Predictive Model to Estimate Learners' Performance through Multisource Data. In *Proceedings of the 2023 2nd International Conference on Futuristic Technologies (INCOFT)*, Belagavi, India, 24–26 November 2023.
16. Harvey, J.L.; Kumar, S.A.P. A Practical Model for Educators to Predict Student Performance in K-12 Education Using Machine Learning. In *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Xiamen, China, 6–9 December 2019.
17. Kaplan, D.; Huang, M. Bayesian Probabilistic Forecasting with Large-Scale Educational Trend Data: A Case Study Using NAEP. *Large-Scale Assess. Educ.* **2021**, *9*, 15.
18. Islam, S. College Student Placement Factors Dataset. *Kaggle* **2021**. Available online: https://www.kaggle.com/datasets/sahilislam007/college-student-placement-factors-dataset (accessed on 12 April 2025).
19. Sekeroglu, B.; Abiyev, R.; Ilhan, A.; Arslan, M.; Idoko, J.B. Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Appl. Sci.* **2021**, 11, 10907.
20. Almusallam, N.; Al Nuaimi, B.T.; Alkattan, H.; Abotaleb, M.; Dhoska, K. Physics-Informed Neural Networks for Solving Heat Equation in Thermal Engineering. *Int. J. Tech. Phys. Probl. Eng.* **2025**, 17, 375–382.
21. Alkattan, H.; Subhi, A.A.; Farhan, L.; Al-mashhadani, G. Hybrid Model for Forecasting Temperature in Khartoum Based on CRU Data. *Mesopotamian J. Big Data* **2024**, *2024*, 164–174.
22. Alkattan, H.; Abbas, N.R.; Adelaja, O.A.; Abotaleb, M.; Ali, G. Data Mining Utilizing Various Leveled Clustering Procedures on the Position of Workers in a Data Innovation Firm. *Mesopotamian J. Comput. Sci.* **2024**, 2024, 104–109.

23. Al-Mahdawi, H.K.; Khorsheed, F.H.; Alhumaima, A.S.; Ramadhan, A.J.; Hussien, K.M.; Alkattan, H. Intelligent Particle Swarm Optimization Method for Parameter Selecting in Regularization Method for Integral Equation. *BIO Web Conf.* **2024**, 97, 00039.
24. Ramadhan, A.J.; Krishna Priya, S.R.; Naranammal, N.; Suman; Lal, P.; Mishra, P.; Abotaleb, M.; Alkattan, H. Use of Random Forest Regression Model for Forecasting Food and Commercial Crops of India. *BIO Web Conf.* **2024**, 97, 00130.
25. Ramadhan, A.J.; Yadav, S.; Anand, S.; Singh, A.P.; Atta, K.; Abotaleb, M.; Alkattan, H.; Albadran, Z. Assessment of Municipal and Industrial Wastewater Impact on Yamuna River Water Quality in Delhi. *BIO Web Conf.* **2024**, *97*, 00124.
26. Ramadhan, A.J.; Krishna Priya, S.R.; Naranammal, N.; Gautam, R.; Mishra, P.; Ray, S.; Abotaleb, M.; Alkattan, H.; Albadran, Z. Use of Factor Scores in Multiple Regression Model for Predicting Customer Satisfaction in Online Shopping. *BIO Web Conf.* **2024**, 97, 00145.
27. Alkattan, H.; Abdulkhaleq Noaman, S.; Subhi Alhumaima, A.; Al-Mahdawi, H.K.; Abotaleb, M.; Mijwil, M.M. A Fusion-Based Machine Learning Framework for Lung Cancer Survival Prediction Using Clinical and Lifestyle Data. *J. Trans. Syst. Eng.* **2025**, 3(2), 382–402.
28. Biswas, J.; Mijwil, M.; Farhan, L.; Alkattan, H. Quantum Computing for Risk Management and Supply Chain Resilience. In *Quantum Computing Applications for Industry 5.0*; CRC Press: Boca Raton, FL, USA, 2024; Chapter 8.
29. Majumder, S.; Alkattan, H.; Rabiei Dastjerdi, H. Quantum Blockchain Scheme for Fraud Detection, Risk Management, and Resilience. In *Quantum Computing Applications for Industry 5.0*; CRC Press: Boca Raton, FL, USA, 2024; Chapter 20.
30. Alkattan, H.; Alhumaima, A.S.; Mohmood, M.S.; AL-Thabhawee, G.D.M. Optimizing Decision Tree Classifiers for Healthcare Predictions: A Comparative Analysis of Model Depth, Pruning, and Performance. *Mesopotamian J. Artif. Intell. Healthc.* **2025**, 3, 124–135.