

Predicting Student Placement Readiness: A Machine Learning Approach Using Coding Activities and Multi-Dimensional Performance Indicators

Priyanshu Chouhan^{1*}, Swapnil Borkar², Sonu Gupta³

Abstract

In the modern information-driven academic world, identifying student employability and placement preparedness has predicted. be made a part and parcel of academic planning and career. development. This study provides a machine learning-based. structure to evaluate and forecast student placement pre-paredness by combining various performance aspects-academic achievement, coding activity, aptitude and behavioral engagement metrics. Multi-source was gathered and preprocessed in the study. student information, such as student records (CGPA, attendance), logging time spent and difficulty of the problem solved, and indicators of co-curricular involvement. Advanced preprocessing data generation, feature engineering and techniques. The dataset quality and volume were improved by using (CTGAN)s. Various machine learning hypothesis, such as Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and XGBoost, were implemented and evaluated. These findings showed that the stacking ensemble model (XGBoost, Random Forest, and Gradient Boosting) was the best and had the highest score having F1-score of 0.7044 and cross-validation consistency (± 0.0253). SHAP interpretability it was found that the most influential are co-curricular involvement, academic consistency, and technical coding proficiency. selection of placement forecasts. The suggested framework offers a predictable instrument that can be effectively interpreted and is clear. Educational institutions may use to detect at-risk students, track whole-child development, and promote data-driven programs of readiness to place.

Keywords: Placement readiness prediction, educational data mining, coding activity analytics, multi-dimensional performance indicators, ensemble learning, XGBoost, SHAP interpretability, co-curricular engagement, predictive analytics in education

INTRODUCTION

Background and Motivation

The world has grown very keen on employability and placement preparedness issues in higher learning institutions. The fast changes in the industrial demands particularly in the field that is more technology-driven demands graduates who are well grounded in academics, have strong practical coding skills, problem solving skills and engage in co-curricular activities. Having been mainly based on CGPA and attendance measures, traditional systems of prediction of placements do not sufficiently reflect the multi-dimensionality of student preparedness to professional settings [1].

The future of machine learning (ML) and educational data mining allows institutions to use heterogeneous information sources to achieve accurate and thorough placement prediction. Through behavioral analytics, coding platform

*Author for Correspondence

Priyanshu Chouhan

E-mail: priynshuchouhn@gmail.com

Research Scholar^{1,2}, Assistant Professor³, Department of Computer Science, Thakur Institute of Management Studies, Career Development & Research, Mumbai, India

Received Date: April 28, 2026

Accepted Date: June 30, 2026

Published Date: July 01, 2026

Citation: Priyanshu Chouhan, Swapnil Borkar, Sonu Gupta. Predicting Student Placement Readiness: A Machine Learning Approach Using Coding Activities and Multi-Dimensional Performance Indicators. Journal of Education Sciences. 2026; 3(2): 29–41p.

performance; engagement metrics combined with academic performance, institutions can design multi-faceted predictive models that are more in line with the modern industry demands [2].

Problem Statement

Although academic success is a fundamental aspect of employability, it may not entirely reflect students' readiness for placement opportunities. Students with similar academic records frequently exhibit considerable variation in coding proficiency, debugging abilities, project experience, and involvement in co-curricular activities. Contemporary predictive algorithms often disproportionately prioritize academic scores or neglect to integrate temporal advancements and behavioral metrics. This creates a critical need for a unified prediction framework that integrates academic, technical, and extracurricular data to provide credible and actionable assessments of placement readiness [3].

Research Objectives

The primary objectives of this study are:

1. To gather and pre-process multi-source data, such as coding logs, attendance records, CGPA, and measures of co-curricular engagement across the various institutional and platform data.
2. To design and test a full suite of coding proficiency characteristics (e.g., proficiency in algorithms, patterns of problem-solving).
3. To compare predictive power of academic, coding and co-curricular features with the cross-validated machine learning models.
4. To maximize the model performance with the help of sophisticated feature selection methods (SHAP analysis, recursive feature elimination) and hyperparameter optimization.
5. To test the final model on hold-out cohorts and test applicability in the real world as applicable to institutions.

Contributions

The key contributions of this work include:

- A novel multi-dimensional feature engineering framework integrating academic, aptitude, coding, and behavioral indicators.
- Comprehensive evaluation of ensemble learning methods for placement readiness prediction.
- SHAP-based interpretability analysis identifying the most influential predictors of student employability.
- A validated predictive system suitable for institutional deployment with actionable insights for student development programs [4].

RELATED WORK

Past Methodologies of Employability Prediction

Initial studies on student employability prediction put a lot of importance on predicting employability through the use of a statistical analysis of academic performance indicators. Ruparel and Swaminarayan compared various machine learning (KNN, LR, SVM) algorithms with an accomplishment of About 94% accuracy, which underlines the significance of preprocessing methods including scaling and outlier management. Nevertheless, their research was also narrowed on academic aspects and excluded behavioral and co-curricular aspects. Eimer and Bohndick did a systematic review of the literature on 21 models of employability and concluded that there is no general framework and assessing employability is very context specific. Their analysis highlighted the necessity of uniform assessment measures and incorporation of different data sources [5].

Deep Learning and Machine Learning Methodologies

Recent research has investigated improved machine learning and deep learning methods on improved prediction accuracy. Sowmia and Poonkuzhali had used Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) to use in employability prediction, where ANN scored 98.52% accuracy. However, their data was not sufficient in quantity and variety, and no soft skills were factored in the prediction model. The

researchers used the methods of Random Forest, XGBoost, Logistic Regression, and SVM to predict placement in campus in Gupta et al., where XGBoost achieving 92.3% accuracy. Importance of features analysis revealed that GPA, projects, and internships were the most important predictors. Although their model gave good results, it did not include behavioral characteristics and gave biased recall in specific case [6].

Coding Activity and Competitive Programming Analysis

Akib et al. used data of competitive programming in Code forces to predict employability, and the researcher categorised the students into four levels of readiness. Their study proved that extracurricular coding practices are good predictors of professional accomplishment. Non-academic variables like communication skills and working as a team were however not incorporated into the model. Jyothi and Poojitha designed a system of prediction of placements by application of the Logistic Regression, Random Forest, and Decision. Random Forest and XGBoost are the best with a higher accuracy rate of 90 percent and 90 percent respectively. They recognized GPA, internships, and certifications as the significant forecasting factors and implemented an interactive dashboard to allow real-time forecasting. Their study was constrained to a single dataset of a single institution, which casts doubts on generalizability across institutions [7].

Hybrid Models and Ensemble Learning

Ensemble learning techniques have proven to be effective in enhancing the robustness and the accuracy of prediction. Moumen et al. performed an extensive review of machine learning techniques between 2016 and 2021 and found that ensemble models have gained better results in terms of employability prediction activities compared to single models. Nonetheless, massive comparative data remain scarce. Ruparel and Swaminarayan suggested a hybrid ensemble technique that incorporates Logistic Regression, Naive Bayes, Gradient Boosting, LDA, KNN, Random Forest, and SVM to give a better precision, recall, and F1-score. The ensemble successfully counterbalanced shortcomings of single models but was confined to the scholarly characteristics without the use of behavioral or interpersonal skills [8].

Competency-Based and Skill Gap Analysis

Smaldone conducted text mining analysis of job advertisements in the field of data science and found that both hard and soft skills are necessary to be employed. Their results also revealed that there is a great discrepancy between the demands of the industry and graduate competencies. None the-less, they only focused on data science jobs and machine learning predictive validation was not present. Chaibate et al. highlighted competency-based training models in order to match educational programs to the market demands. Their systematic review revealed that there was a gap in the literature on empirical research that would prove the efficacy of such frameworks in enhancing student employability results [9].

Research Gaps

The literature analysis indicates that there are some significant gaps in literature:

- *Weak Multi-Source Integration:* There is not much literature that links coding activity logs with academic and co-curricular data into one framework.
- *Absence of Interpretability:* Not all high-accuracy models have interpretability mechanisms (e.g., SHAP values) that can be used to make decisions in an institution.
- *Weak Stress on At-Risk Students:* Models on the process of recall and early detection of students to be provided with intervention.
- *Lack of Behavioral Analytics:* Soft skills, interaction pattern, and time dynamics have not been well studied in predictive models.
- *Generalizability Concerns:* In the majority of the models, cross-institutional testing is the one that is performed in a small sample, generally on a single-institution dataset.

To fill in these gaps, this study will present a holistic multi-dimensional model, which

combines academic, aptitude, coding, and behavioral indicators with ensemble learning and SHAP-based interpretability analysis [10].

METHODOLOGY

Data Collection and Sources

The dataset used in this study was compiled from multiple institutional and platform-based sources to capture a comprehensive picture of student performance and engagement across various dimensions.

Coding Activity Logs

Coding activity data were extracted from Campus Rank, a competitive programming and learning platform. The collected metrics include:

- Total coding time and problem-solving frequency
- Difficulty-wise problem counts (easy, medium, hard)
- Behavioral engagement metrics: dominant page ratio, competitive focus ratio, page entropy

These metrics quantify not only technical proficiency but also consistency, focus, diversity, and interaction patterns on the platform.

Academic Records

Institutional academic data consists of:

- Attendance percentage
- Secondary School Certificate (SSC) percentage
- Higher Secondary Certificate (HSC) percentage
- Graduation CGPA/percentage
- Master of Computer Applications (MCA) CGPA (converted to percentage using the formula: $MCA_{\%} = 7.1 \times CGPA + 11$)

These academic indicators were aggregated into composite indices representing both performance level and consistency.

Aptitude and Cognitive Performance

Standardized aptitude assessment scores were collected in three categories:

- English language proficiency
- Quantitative reasoning ability
- Logical reasoning capability

Co-curricular and Extra-curricular Involvement

Binary indicators were derived from institutional records and surveys representing student participation in:

- Technical hackathons and coding competitions
- Technical clubs and professional societies
- Workshops, seminars, and training programs
- Paper presentations and research activities
- Leadership roles in student organizations

Data Preprocessing Pipeline

A rigorous data preprocessing pipeline was implemented to ensure data quality, consistency, and suitability for machine learning modeling.

Data Cleaning

- Handled missing and duplicate entries

- Removed null or irrelevant fields
- Standardized categorical variables (e.g., “yes/no” converted to binary 1/0)
- Applied anonymization protocols to protect student identities

Missing Value Imputation

- Numerical columns: imputed using mean values
- Categorical columns: imputed using mode values
- Distribution consistency maintained through stratified imputation

Normalization and Scaling

Features with varying scales were standardized using RobustScaler, which reduces the effect of outliers compared to traditional Min-Max or z-score scaling. RobustScaler uses the median and interquartile range for scaling, making it particularly suitable for educational datasets with potential outliers.

Data Balancing

To address class imbalance between “ready” and “not-ready” students, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances, thereby improving model sensitivity to underrepresented categories.

Data Augmentation

For datasets with limited sample size, Conditional Tabular Generative Adversarial Network (CTGAN) was employed to synthetically generate additional realistic samples. CTGAN models the probability distribution of the original dataset and generates new samples that preserve statistical properties and feature correlations, expanding the dataset to approximately 1000 records for improved model generalization [11].

Feature Engineering

Feature engineering played a crucial role in integrating multi-dimensional attributes of student performance into quantifiable metrics. The following composite features were derived.

Academic Strength

These scores were averaged to form an overall aptitude strength feature, representing general employability readiness.

$$\text{Academic Strength} = 0.2 \times \text{SSC} + 0.3 \times \text{HSC} + 0.25 \times \text{Grad} + 0.25 \times \text{MCA}_{\%} \quad (1)$$

This weighted average represents the overall academic foundation, with higher weights assigned to more recent educational stages.

Aptitude Strength:

$$\text{Aptitude strength} = \frac{\text{English} + \text{Quantitative} + \text{Logical}}{3} \quad (2)$$

This feature captures analytical and reasoning abilities essential for professional success.

Coding Performance Metrics

$$\text{Difficulty Score: Difficulty Score} = 1 \times \text{Easy} + 2 \times \text{Medium} + 3 \times \text{Hard} \quad (3)$$

$$\text{Coding Proficiency: Coding Proficiency} = 0.7 \times \text{Difficulty Score} + 0.3 \times \text{Coding Activity} \quad (4)$$

$$\text{Coding Engagement Index: Engagement Index} = 0.3 \times \text{Dominant Page Ratio} + 0.4 \times \text{Competitive Focus Ratio} + 0.3 \times \text{Page Entropy} \quad (5)$$

$$\text{Total Coding Strength: Total Coding Strength} = 0.7 \times \text{Coding Proficiency} + 0.3 \times \text{Engagement Index} \quad (6)$$

Interaction and Consistency Features

Tech Learning Synergy: Tech Learning Synergy = Total Coding Strength $\times$ Aptitude Strength (7)

This interaction term captures the synergy between technical skills and cognitive abilities.

Academic Consistency: $Academic\ Consistency = Academic\ Strength \times \frac{Attendance}{100}$ (8)

This feature scales academic performance by attendance, representing stability and regularity.

Performance Index: Performance Index = $0.22 \times Academic\ Strength + 0.13 \times Aptitude\ Strength + 0.45 \times Total\ Coding\ Strength + 0.20 \times Tech\ Learning\ Synergy$

This holistic composite feature integrates all dimensions of student readiness with empirically determined weights that reflect their relative importance in employability prediction [12].

Machine Learning Models

Baseline Model

The baseline model was implemented in terms of Logistic Regression so as to determine a model that is interpretable in terms of its performance benchmark. Logistic Regression offers a linear line of decision and acts as the point of reference to assess more complex models

Tree-Based Ensemble Models

- **Random Forest:** An ensemble of decision trees trained on bootstrap samples with random feature selection, providing robust multi-feature decision learning and natural feature importance metrics.
- **Gradient Boosting:** Enhances the tree building process by successively enhancing earlier models by increasing their accuracy, through the use of an ensemble non-linear patterns optimization.
- **XGBoost:** The most advanced implementation of gradient boosting that supports regularization, provides options of missing values and parallel processing. XGBoost was chosen due to its high quality performance in tasks of predicting structured data [13].

Hyperparameter Tuning

Hyperparameters of XGBoost were optimized using RandomizedSearchCV with 10 fold stratified cross-validation:

- Learning rate: [0.01, 0.05, 0.1]
- Maximum tree depth: [4, 6, 8]
- Number of estimators: [200, 400, 600]
- Subsample ratio: [0.7, 0.8, 1.0]
- Column sampling: [0.7, 0.8, 1.0]
- Gamma (minimum split loss): [0, 0.1, 0.2]
- L2 regularization (lambda): [0.5, 1, 1.5]
- L1 regularization (alpha): [0, 0.1, 0.5]

Randomized search was done on 50 combinations of parameters, trade-off between the precision and the recall by maximizing F1-score.

Stacking Ensemble

A meta-learning ensemble was constructed combining:

- **Base learners:** Tuned XGBoost, Random Forest (300 estimators), Gradient Boosting (300 estimators)
- **Meta-classifier:** Logistic Regression with maximum 2000 iterations

- *Cross-validation*: 5-fold stratified CV for base model predictions

The stacking ensemble leverages the complementary strengths of diverse base models, with the meta-classifier learning optimal combinations of base predictions to enhance overall accuracy and generalization.

Model Evaluation Strategy

Training-Test Split

Data were partitioned into 80% training and 20% testing subsets using stratified sampling to maintain class distribution balance. This holistic composite feature integrates all dimensions of student readiness with empirically determined weights that reflect their relative importance in employability prediction [12].

Cross-Validation

10-fold Stratified Cross-Validation was performed to ensure:

- Performance stability across diverse data subsets
- Reliable estimation of model generalization
- Detection of overfitting or underfitting
- Statistical significance of performance differences

Evaluation Metrics

Models were evaluated using multiple complementary metrics:

- *Accuracy*: Overall correctness of predictions
- *Precision*: Proportion of predicted “ready” students actually ready
- *Recall*: Proportion of actual “ready” students correctly identified
- *F1-Score*: Harmonic mean of precision and recall, particularly suitable for imbalanced datasets
- *ROC-AUC*: Classification performance independent of decision thresholds

F1-score was selected as the primary optimization metric due to the importance of balancing false positives (incorrectly predicting readiness) and false negatives (missing at-risk students).

Interpretability Analysis

SHAP (Shapley Additive explanations) values were computed for post-hoc model in-terpretability. SHAP provides:

- Individual feature contribution to each prediction
- Global feature importance rankings
- Directional impact (positive/negative) of features
- Consistency with game-theoretic fairness principles

SHAP analysis enables identification of the most influential academic, aptitude, coding, and behavioral factors affecting placement readiness predictions, providing actionable insights for institutional intervention strategies [14].

RESULTS AND DISCUSSION

Model Performance Comparison

Multiple machine learning models were trained and evaluated on the placement readiness dataset. Table I presents the comprehensive performance comparison across all implemented models.

The Stacking Ensemble model achieved the best overall performance with an F1-score of 0.7044, demonstrating superior balance between precision and recall. Random Forest showed the highest individual model F1-score (0.744), but the stacking approach improved generalization by combining diverse model predictions (Table 1).

Cross-Validation Results

10-fold Stratified Cross-Validation was performed on the stacking ensemble model to assess stability and generalization. The cross-validation yielded an average F1-score of 0.7058 ± 0.0253 , indicating:

- Strong consistency across different data partitions
- Low variance suggesting stable predictions
- Effective generalization to unseen data
- Minimal overfitting to training data

The small standard deviation (± 0.0253) confirms that model performance is reliable and not dependent on specific train-test splits.

Confusion Matrix Analysis

The confusion matrix for the best-performing stacking ensemble model revealed:

- *True Negatives*: 183 students correctly identified as not-ready
- *True Positives*: 168 students correctly identified as ready
- *False Positives*: 63 students incorrectly predicted as ready
- *False Negatives*: 78 students incorrectly predicted as not-ready

The balanced distribution of correct predictions across both classes demonstrates that the model effectively handles the inherent class imbalance through SMOTE preprocessing and ensemble learning. The reduction in false negatives compared to individual models (e.g., XGBoost had 89 false negatives) indicates improved sensitivity in identifying students who are actually ready for placement (Figure 1).

Feature Importance Analysis

SHAP (SHapley Additive explanations) analysis was performed to interpret model behavior and identify the most influential predictors of placement readiness. Table II presents the ranking of features based on their impact on model predictions [15].

The SHAP analysis reveals several critical insights:

Co-curricular Involvement

Co-curricular participation emerged as the single most influential predictor of placement readiness. Students actively engaged in technical competitions, hackathons, workshops, and leadership activities demonstrated significantly higher placement probability (Figure 2).

This finding aligns with industry expectations that real-world exposure, teamwork, communication skills, and practical problem-solving are essential employability competencies beyond academic achievement (Table 2).

- *Academic Consistency*: Academic consistency (academic strength scaled by attendance) ranked second in importance, indicating that sustained effort and regular engagement are more predictive than isolated academic performance. Recruiters often view reliability, discipline, and continuous improvement as key professional traits, which academic consistency effectively captures.
- *Technical Proficiency*: Total coding strength, combining coding proficiency with behavioral engagement patterns, contributed significantly to predictions. This validates the hypothesis that sustained coding activity and diverse problem-solving experiences rather than sporadic achievements reflect genuine technical readiness.
- *Holistic Performance Index*: The performance index and tech learning synergy features captured multi-domain learning and integration of aptitude with technical skills. Their moderate but consistent SHAP influence supports the value of composite indicators in employability prediction systems [19].

Table 1. Performance comparison of machine learning models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.565	0.578	0.562	0.570
Random Forest	0.727	0.716	0.774	0.744
SVM	0.712	0.727	0.701	0.714
Gradient Boosting	0.685	0.685	0.715	0.700
XGBoost (Tuned)	0.675	0.689	0.638	0.662
Stacking Ensemble	0.713	0.727	0.683	0.704

Discussion of Findings

- *Multi-Dimensional Integration:* The findings indicate that multi-domain employability predictions will provide more accurate predictions when based on academic, aptitude, coding, and behavioral signals as compared to uni-domain models. The better results of the stacking ensemble (F1 = 0.704) versus academic-only baselines (F1 = 0.570) the argument that combined models are better than single-category methods are supported [16].
- *Behavioral Factors:* Co-curricular participation and academic consistency is dominant in SHAP analysis, which proves the significance of the behavior consistency. Such findings indicate that soft skills, patterns of engagement, and holistic development are essential aspects, which are usually not accounted in conventional academic-contained prediction systems [17].
- *Technical Depth:* Although in SHAP importance, technical coding strength was the seventh most important part of the entire predictive framework, it was still a vital part of the entire predictive system. The interaction between proficiency and aptitude in coding (technical learning synergy) shows that technical proficiency is best applied in combination with high levels of analytical underpinnings.
- *Ensemble Learning Effectiveness:* The effectiveness of the stacking ensemble in terms of generalization and lower variance proves the effectiveness of ensemble combination of different base learners. The meta-classifier was able to learn the optimum weightings of base model predictions which are more precise and recalling than single models [18].

Comparative Analysis with Prior Work

compares the proposed model with relevant prior studies in student employability prediction.

While direct accuracy comparison is challenging due to dataset differences, the proposed model achieves competitive performance with enhanced interpretability through SHAP analysis and richer feature diversity spanning academic, technical, and behavioral dimensions (Table 3).

Practical Implications

The findings have several practical implications for educational institutions:

- *Early Intervention:* The model can identify at-risk students early in their academic journey, enabling timely career counseling and skill development programs.
- *Holistic Development:* The importance of co-curricular participation justifies institutional investment in hackathons, technical clubs, and leadership development programs [20].
- *Personalized Guidance:* SHAP-based individual predictions provide actionable feedback on specific areas requiring improvement (e.g., coding practice, attendance, co-curricular engagement) [21].
- *Curriculum Design:* The synergy between technical skills and aptitude suggests value in integrated learning approaches combining coding practice with analytical reasoning development [22].
- *Placement Strategy:* Placement officers can prioritize resources toward students with lower predicted readiness, maximizing overall placement outcomes [23].

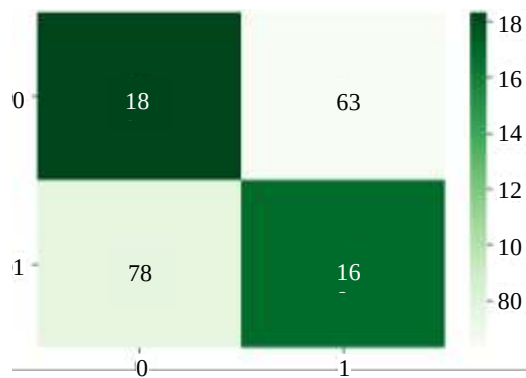


Figure 1. Confusion Matrix – Tuned XGBoost.

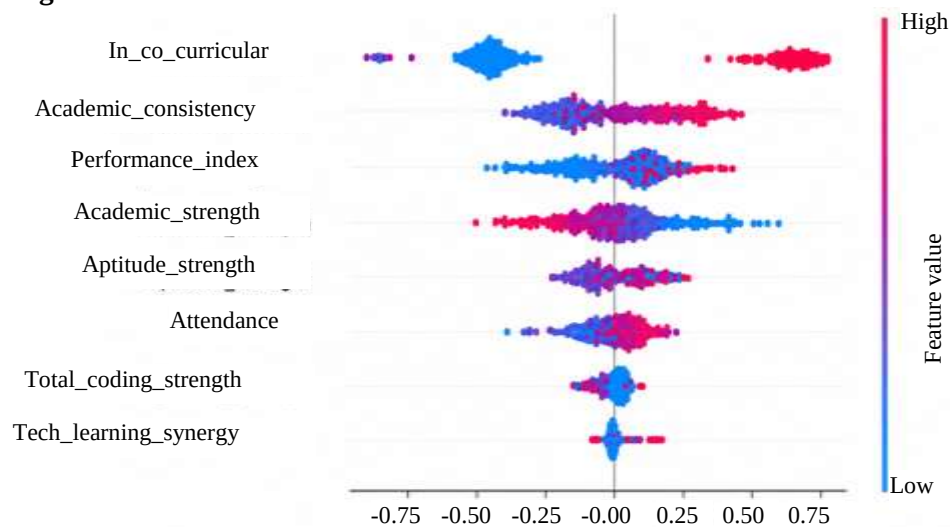


Figure 2. SHAP Summary plot for feature importance.

Table 2. SHAP-based feature importance rankings.

Rank	Feature	Impact Description
1	Co-curricular Participation	Highest positive influence on readiness
2	Academic Consistency	Strong positive correlation with placement success
3	Performance Index	High overall influence combining multiple dimensions
4	Academic Strength	Moderate positive influence
5	Aptitude Strength	Moderate positive influence on analytical readiness
6	Attendance	Moderate positive correlation with consistency
7	Total Coding Strength	Moderate but essential technical indicator
8	Tech Learning Synergy	Mild positive interaction effect

Table 3. Comparison with prior research.

Study	Model	Features	Performance
Ruparel et al.	KNN, SVM	Academic	94% Acc
Gupta et al.	XGBoost	Academic + Behavioral	92.3% Acc
Jyothi et al.	RF, XGBoost	Academic + Soft Skills	90% Acc
Proposed	Stacking	Multi-dimensional	0.704 F1

CONCLUSION

This study introduced an all-encompassing machine learning model to predict student placement readiness integrating academic performance, aptitude testing, coding activity measurements and

behavioral engagements measures. The suggested system overcomes major constraints of the current academic-based prediction models, which include multi-incorporation. Dimensional indicators are indicators that are more representative of the industry needs today. The stacking ensemble, which was a XGBoost, Random Forest and Gradient Boosting with a Logistic Regression meta-classifier, had the highest performance with F1-score of 0.7044 and cross-validation consistency of 0.7058 $\pm$ 0.0253. This was better than single models and was shown to be high in generalization with a variety of student profiles. SHAP-based interpretability analysis showed that the most predictive variables of placement readiness are co-curricular participation, academic consistency, and performance index, which validates the need to consider the whole types of the holistic development of students rather than the academic success only. The framework does offer a clear, practical information, which can be used by educational institutions to perform early intervention, personalized guidance, and optimization of placement strategies based on data.

Key Findings

The contributions and findings of the research are the following:

1. *Dominance*: Co-curricular activities and academic performance were the best indicators, showing how important soft skills and patterns of engagement are for getting a job.
2. *Technical-analytical synergy*: The connection between coding competency and aptitude strength demonstrates the significance of skill integration.
3. *Ensemble multi-dimensional integration*: Using academic, aptitude, coding, and behavioural features makes forecasts far more accurate and reliable than using only one type of feature.
4. *Behavioral superiority*: Stacking ensemble learning works better at generalising and has less variation in predictions since it combines several base models in a successful way.
5. *Interpretable AI*: SHAP-based feature importance analysis supports making decisions and taking actions related to the institution.

Limitations

Despite achieving satisfactory performance, this study has several limitations:

1. *Dataset size*: The original dataset itself was relatively small and thus synthetic augmentation had to be carried out using CTGAN. Although augmentation enhanced model generalization, it should be tested on bigger and real-world datasets.
2. *Missing behavioral data*: Certain behavioral measures (platform usage logs, engagement patterns) had sparse or incomplete data, which restricted the ability to do a time analysis.
3. *Static feature assumption*: The model considers the student attributes to be static rather than dynamic because employability readiness is dynamic and changes over time.
4. *Low level of soft skill coverage*: There was no coverage of psychological aspect, communication skills, emotional intelligence and performance in an interview because the data are not available.
5. *Single-institution test*: This model was created and tested by means of a single institution, which casts doubt upon cross-institutional extrapolation. could analyze student portfolios, resumes, and project descriptions for additional employability signals.
6. *Cross-institutional validation*: Check the model with data of several colleges and universities in order to evaluate the generalizability of the model to another academic context, curriculum, and student population. Transfer learning strategies would be able to adapt the models on one institution to other institutions with smaller data.
7. *Causal analysis*: Go beyond correlation based pre-diction to causal inference models that can determine the interventions that are most likely to enhance student employability. Methods, like propensity score matching or instrumental variable analysis, would be able to furnish a causal relationship between particular activities and outcomes of placement.
8. *GitHub and portfolio analysis*: Incorporate NLP-based analysis of student GitHub repositories, project portfolios, and technical blogs to elicit other indicators of technical competence, collaboration abilities, and life-long learning practices.

Concluding Remarks

This study proves that to determine the readiness of students to be placed, one needs to look at the whole picture that involves academic excellence, technical mastery, analytical thinking, and behavioral interaction. The suggested machine learning model with ensemble learning and SHAP interpretability is a powerful and clear-cut resource that can guide learning institutions in identifying vulnerable students, streamlining placement practices, and enabling data-driven career building models. The fact that this paper finally goes beyond the old-fashioned academic-centric models and represents the overall multi-dimensional prediction helps to further strengthen the emerging branch of intelligent educational analytics, and offer a basis to future studies in the optimization of student success with AI.

REFERENCES

1. Adeniyi J, Adeniyi AE, Yetunde JO, Egbedokun GO, Ajagbe KD, Obuzor PC, Ajagbe SA. Comparative analysis of machine learning techniques for the prediction of employee performance. *ParadigmPlus*. 2022;3(3):1-5.
2. Eimer A, Bohndick C. Employability models for higher education: A systematic literature review and analysis. *Social Sciences & Humanities Open*. 2023 Jan 1;8(1):100588.
3. Singh K. Exploring Artificial Intelligence: A Deep Review of Foundational Theories, Applications, and Future Trends. *Journal of Scientific Innovation and Advanced Research (JSIAR)*. 2025 Sep;1(6):295-305.
4. Ahmad N, Ghadi Y, Adnan M, Ali M. Load forecasting techniques for power system: Research challenges and survey. *IEEE access*. 2022 Jul 1;10:71054-90.
5. Asmara RA, Rosiani UD, Mentari M, Syulistyo AR, Shoumi MN, Astiningrum M. An experimental study on deep learning technique implemented on low specification openmv cam h7 device. *JOIV: International Journal on Informatics Visualization*. 2024 May 31;8(2):1017-29.
6. Maphisa X, Nkadimeng M, Telukdarie A. Contextual Intelligence: An AI Approach to Manufacturing Skills' Forecasting. *Big data and cognitive computing*. 2024 Sep 2;8(9):101.
7. Baranyi V. Systematic literature review on the digital transformation of the personnel selection process. *German Journal of Human Resource Management*. 2026 May;40(2):223-54.
8. Selvakani S, Vasumathi K. Enhancing Dropout Prediction in Higher Education using a Hybrid Machine Learning Approach. *Journal of Computer Science (ISSN NO: 1549-3636)*. 2025 May;18(05).
9. Chen Y, Zhang J, Li Z, Zhang H, Chen J, Yang W, Yu T, Liu W, Li Y. Manufacturing technology of lightweight fiber-reinforced composite structures in aerospace: Current situation and toward intellectualization. *Aerospace*. 2023 Feb 22;10(3):206.
10. Becker BE, Huselid MA. Strategic human resources management: where do we go from here?. *Journal of management*. 2006 Dec;32(6):898-925.
11. Ruparel M, Swaminarayan P. Enhanced student placement prediction using machine learning: A comparative evaluation of algorithms. *International Journal of Engineering Trends and Technology*. 2025;73(1):225-36.
12. Jackson D, Dean BA. Employability-related activities beyond the curriculum: how participation and impact vary across diverse student cohorts. *Higher Education*. 2023 Nov;86(5):1151-72.
13. Drogkoula M, Kokkinos K, Samaras N. A comprehensive survey of machine learning methodologies with emphasis in water resources management. *Applied Sciences*. 2023 Nov 8;13(22):12147.
14. Córdova-Esparza DM, Terven J, Romero-González JA, Córdova-Esparza KE, López-Martínez RE, García-Ramírez T, Chaparro-Sánchez R. Predicting and preventing school dropout with business intelligence: Insights from a systematic review. *Information*. 2025 Apr 19;16(4):326.
15. Mustaqeem M, Alam M, Mustajab S, Alshanketi F, Alam S, Shuaib M. Comprehensive bibliographic survey and forward-looking recommendations for software defect prediction: datasets, validation methodologies, prediction approaches, and tools. *IEEE Access*. 2024 Dec 13;13:866-903.
16. Sridharan R, Sukanya NS, Sivasakthi B, Sankaran V. Machine learning/deep learning use cases in

- education. In *Artificial Intelligence based Solutions for Industrial Applications* 2024 Nov 20 (pp. 124-148). CRC Press.
17. Muhammad HB, Muhammad AM, Abdullahi SD. Machine learning for predicting students' employability. *UMYU Scientifica*. 2023 Mar 30;2(1):1-9.
 18. Liu C, Wang GC, Wang HF. The application of artificial intelligence in engineering education: A systematic review. *IEEE Access*. 2025 Jan 23;13:17895-910.
 19. Tang Y, Moretti R, Meiler J. Recent advances in automated structure-based de novo drug design. *Journal of Chemical Information and Modeling*. 2024 Mar 14;64(6):1794-805.
 20. Xing Z, Wu X, Li J. Student Employment Forecasting Model Based On Random Forest and Multi-features Fusion [J]. *Journal of Applied Science and Engineering*. 2025;29(2):329-36.
 21. Yu H, Mu Q, Liu C, Wang S, Sun J. Technical system of electroencephalography-based brain-computer interface: Advances, applications, and challenges. *Neural Regeneration Research*. 2026 Sep 1;21(9):3885-907.
 22. Stošić L, Radonjić A, Krčadinac O, Baltezarević B, Mikhailova O. Personalized Learning through Artificial Intelligence: Opportunities, Risks, and Policy Perspectives. *International Journal of Cognitive Research in Science, Engineering and Education*. 2025;13(2):541-9.
 23. Kilar W, Brzosko-Sermak A. Funkcjonowanie przedsiębiorstw w wybranych krajach. *Studies of the Industrial Geography Commission of the Polish Geographical Society*. 2024 Jun 27;38(2).