

## SMART PLACEMENT PREDICTION SYSTEM: LEVERAGING MACHINE LEARNING FOR CAREER OUTCOME FORECASTING

**Gagandeep Singh Kainth**

Student, GNA University, Phagwara

**Dr. Rajesh Sharma**

Associate Professor, GNA University, Phagwara

### Abstract

*Student placement forecasting is a very important problem in institutions of higher learning that may be employed to make predictive interventions and effective resource allocation to enhance employment. Conventional techniques of placement assessment are highly reliant on manual analysis and historical patterns, which are labour intensive, subjective, and may be ineffective in dealing with complicated associations among academic, technical, and interpersonal abilities. The present paper shows PlacePredict, an intelligent prediction framework designed for forecasting the success of student placements using a significant collection of features comprising academic achievements, technical skills, experiential learning, and soft skills. Five supervised machine learning algorithms were trained and evaluated using a publicly available Kaggle dataset. A separate institutional dataset collected from GNA University was subsequently used to demonstrate the applicability of the proposed prediction framework to real-world student records. A SMOTE-based class balancing approach was applied to balance minority class samples. Experimental findings indicate that the SVM-based classifier has the strongest overall performance. The suggested framework offers practical information to educational institutions to detect at-risk students and develop specific intervention plans.*

**Keywords:** Student Placement Prediction, Machine Learning, SVM, SMOTE, EDM.

### Introduction

#### 1.1 Background

The transition from academic to professional employment is a milestone in the academic and professional career of students. In contemporary learning institutions, the office of campus placements

has a major role in linking students with the job market. The result of any placement is however vastly different because of various factors including academic results, technical knowledge, experience and interpersonal skills.

Student placement prediction is a supervised learning problem in which student past data are utilized to predict future placement outcomes (Yağci et al., 2022; Varaprasad et al., 2025). The current progress in machine learning and access to student data have made predictive models possible that assist institutions to identify at-risk students and take preventive measures on time (Cortes & Vapnik, 1995; Bishop, 2006).

### ***1.2 Shortcomings of Traditional Methods***

The traditional placement assessment is based on manual assessment, historical tendency and subjectivity (Han et al., 2011). These techniques are time-consuming and do not capture complex non-linear relationships among multiple variables (Bishop, 2006).

### ***1.3 Research Gap and Contribution***

Existing studies on student placement prediction primarily focus on improving predictive accuracy while providing limited details regarding validation methodology, class imbalance handling, and reproducibility. Furthermore, many studies evaluate their models using only publicly available datasets without demonstrating their applicability to real institutional student data.

This study presents PlacePredict, a machine learning framework trained and evaluated on a publicly available Kaggle dataset using multiple classification algorithms. To assess its practical applicability, the trained framework was subsequently applied to institutional student records collected from GNA University. This approach demonstrates how predictive models developed using publicly available educational datasets can support institutional decision-making while maintaining methodological transparency.

## **Related Work**

### ***2.1 Literature Review***

The prediction of student placement with different machine learning methods has been widely researched using approaches ranging from traditional models to ensemble techniques (Gupta et al., 2025; Kuhn & Johnson, 2013).

**Table 1: Comparative Analysis of the Recent Studies on Student Placement Prediction.**

| Author and Year                     | Machine Learning Method(s)                                                                                             | Key Findings and Performance                                                                                                                                 | Limitations                                                                      |
|-------------------------------------|------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Mustafa Yağci et al., 2022          | Random Forest, kNN, Support Vector Machine, Logistic Model, Bayesian Classifier                                        | Random Forest produced the highest accuracy of 0.985; SVM achieved 0.845; kNN achieved 0.818; Naïve Bayes achieved 0.815; Logistic Regression achieved 0.701 | Insufficient evidence of validation methodology and generalizability assessment  |
| Shiv Kumar Tiwari, 2025             | Decision Tree, Random Forest, SVM, Gradient Boosting combined in hybrid ensemble; Explainable AI applied               | Hybrid ensemble reported overall accuracy of 89% with model interpretation via explainable AI methods                                                        | Insufficient evidence of cross-validation and external validation procedures     |
| Kurapati V. Varaprasad et al., 2025 | Logistic Model, Tree-Based Classifier, Random Forest Approach, K-Nearest Neighbors, Bayesian Method                    | Reported accuracies: Logistic Regression 76.92%; Decision Tree 96.92%; Random Forest 89.23%; KNN 96.92%; Naïve Bayes 100%                                    | Insufficient evidence of class imbalance handling and metric selection rationale |
| Sachin Bhoite et al., 2023          | Logistic Model, Support Vector Classifier, K-Nearest Neighbors, Tree Classifier, Forest-Based Model, Adaptive Boosting | Framed probabilistic ensemble approach with multiple classifiers; performance numbers not provided in abstract                                               | Insufficient evidence of detailed performance metrics and comparative analysis   |
| Sumedh Tardalkar et al., 2024       | Logistic Model, Decision-Based Tree Algorithm, Random Forest Method, SVM Classifier                                    | Random Forest reported best performance with accuracy approximately 81.47% on dataset of 2966 records                                                        | Insufficient evidence of feature engineering and preprocessing strategies        |
|                                     | Forest Method, SVM Classifier                                                                                          |                                                                                                                                                              |                                                                                  |

|                           |                                                                    |                                                                                                                           |                                                                          |
|---------------------------|--------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Shilpa Gupta et al., 2025 | Random Forest, XGBoost, Regularized Logistic Regression, SVM (RBF) | Ensemble techniques outperformed traditional algorithms; XGBoost reported 92.3% accuracy with feature importance analysis | Insufficient evidence of hyperparameter tuning and validation procedures |
|---------------------------|--------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|

## Preprocessing and Visualization of Data

### 3.1 Data Source

Two independent datasets were used in this study for different purposes. The machine learning models were developed and evaluated using the publicly available Engineering Student Journey dataset obtained from Kaggle, which contained 2,000 student records with academic, technical, experiential, and placement-related attributes. The Kaggle dataset included 1,032 placed students and 967 not placed students, indicating a slight class imbalance prior to preprocessing.

In addition, a primary institutional dataset consisting of 32 student records was collected from GNA University through a structured questionnaire. This dataset was not merged with the Kaggle dataset for model training. Instead, it was used to demonstrate the applicability of the trained prediction framework to institutional student data. Since both datasets shared similar feature definitions, the same preprocessing pipeline was applied to the institutional records before generating placement predictions. (<https://www.kaggle.com/datasets/rakeshkapilavai/engineering-student-journey>)

### 3.2 Feature Description

The predictor feature set is sorted in four conceptual categories of various dimensions of student preparedness and capability.

**The entire features that are encompassed in the model have been well described in table3.**

**Table 3: Description of Features**

| Feature Category            | Feature Name | Description                    | Data Type  | Range/Values |
|-----------------------------|--------------|--------------------------------|------------|--------------|
| <i>Academic Performance</i> | CGPA         | Cumulative Grade Point Average | Continuous | 1.0 - 10.0   |

|  |                 |                                            |            |                 |
|--|-----------------|--------------------------------------------|------------|-----------------|
|  | 10th Percentage | 10th standard examination score            | Continuous | 0 - 100         |
|  | 12th Percentage | 12th standard or diploma examination score | Continuous | 0 - 100         |
|  | Active Backlogs | Presence of pending course backlogs        | Binary     | 0 (No), 1 (Yes) |

|                              |                          |                                            |          |                                              |
|------------------------------|--------------------------|--------------------------------------------|----------|----------------------------------------------|
| <b>Technical Skills</b>      | Coding Skill Level       | Self-assessed programming proficiency      | Ordinal  | 1 (Beginner), 2 (Intermediate), 3 (Advanced) |
|                              | DSA Knowledge            | Data Structures and Algorithms proficiency | Ordinal  | 1 (Basic), 2 (Medium), 3 (High)              |
|                              | Programming Languages    | Number of programming languages known      | Discrete | 1 - 8                                        |
|                              | Technical Certifications | Possession of technical certifications     | Binary   | 0 (No), 1 (Yes)                              |
| <b>Experiential Learning</b> | Number of Internships    | Count of completed internships             | Discrete | 0 - 5+                                       |
|                              | Number of Projects       | Count of completed technical projects      | Discrete | 0 - 10+                                      |
|                              | Hackathon Participation  | Participation in coding competitions       | Binary   | 0 (No), 1 (Yes)                              |
| <b>Soft Skills</b>           | Communication Skills     | Self-assessed communication ability        | Ordinal  | 1 - 5                                        |
|                              | Interview Confidence     | Self-assessed interview confidence         | Ordinal  | 1 - 5                                        |
|                              | Logical Reasoning        | Logical reasoning ability                  | Ordinal  | 1 (Low), 2 (Medium), 3 (High)                |
| <b>Behavioral</b>            | Willingness to Improve   | Attitude toward skill development          | Ordinal  | 1 - 5                                        |

## Data Preprocessing

Data pre-processing involved handling missing values, encoding categorical variables, and standardizing feature formats prior to model development. Numerical features were scaled where required to ensure compatibility with machine learning algorithms. The original Kaggle dataset exhibited a slight class imbalance between placed and non-placed students. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training data during model development. The test data remained completely unseen throughout the training process to prevent information leakage and ensure an unbiased evaluation of model performance.

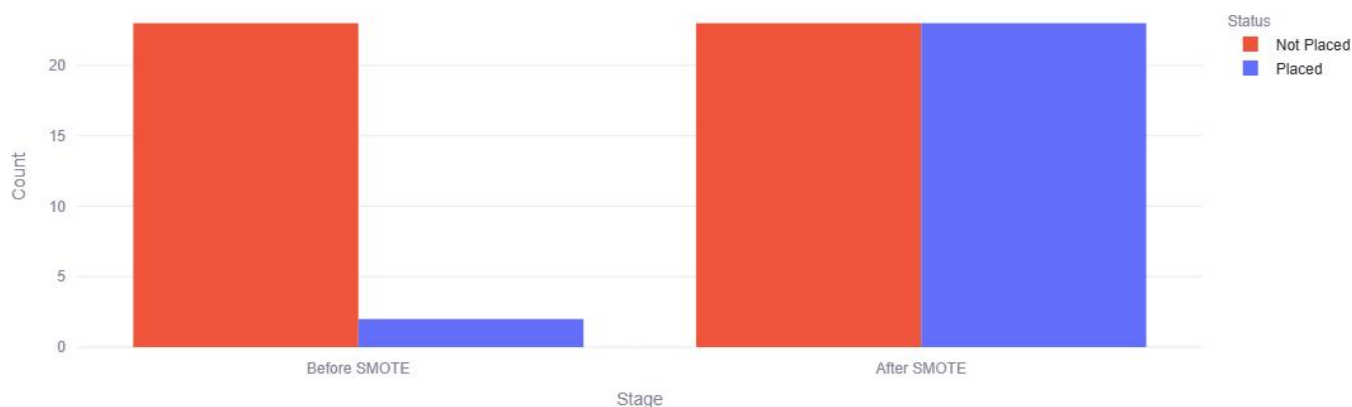


**Figure 1: Class balance comparison before and after applying SMOTE**

As shown in Figure 1, the dataset was initially imbalanced, with a majority of placed students. After applying SMOTE, the classes were balanced, improving the model's ability to learn both classes effectively.

## Methodology

### *Problem Formulation*

Student placement prediction is formulated as a supervised binary classification problem. The target variable, **Placement Status**, is defined as:

**1 (Placed):** The student successfully secured employment through the campus placement process.

**0 (Not Placed):** The student did not obtain employment through the campus placement process.

The objective of the predictive models is to classify each student into one of these two classes based on academic performance, technical skills, experiential learning, and soft-skill related features.

### *Machine Learning Pipeline*

The proposed prediction framework followed the following workflow:

1. Data preprocessing and cleaning.
2. Feature encoding and numerical scaling.

3. Stratified partitioning of the dataset into training and testing subsets.
4. Application of SMOTE only on the training dataset to address class imbalance.
5. Training of multiple machine learning classifiers.
6. Hyperparameter optimization for each classifier.
7. Model evaluation using stratified cross-validation.
8. Selection of the best-performing model based on multiple evaluation metrics.

#### ***4.1 Experimental Setup***

The machine learning models were developed using the Kaggle Engineering Student Journey dataset. The dataset was partitioned into **80% training data and 20% testing data** using a **stratified train-test split** with **random\_state = 42** and data shuffling enabled. Stratification ensured that the proportion of placed and non-placed students was preserved in both subsets. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied **only to the training data** using **random\_state = 42** and **k\_neighbors = 1**, thereby preventing information leakage into the test set. Numerical features were standardized using **StandardScaler**, where the scaler was fitted on the training data and subsequently applied to the test data.

### **Results and Discussion**

#### ***5.1 Model Comparison***

Here is the, full comparison of all the five classification algorithms in all the evaluation measures which gives an overall performance picture of the relative performance.

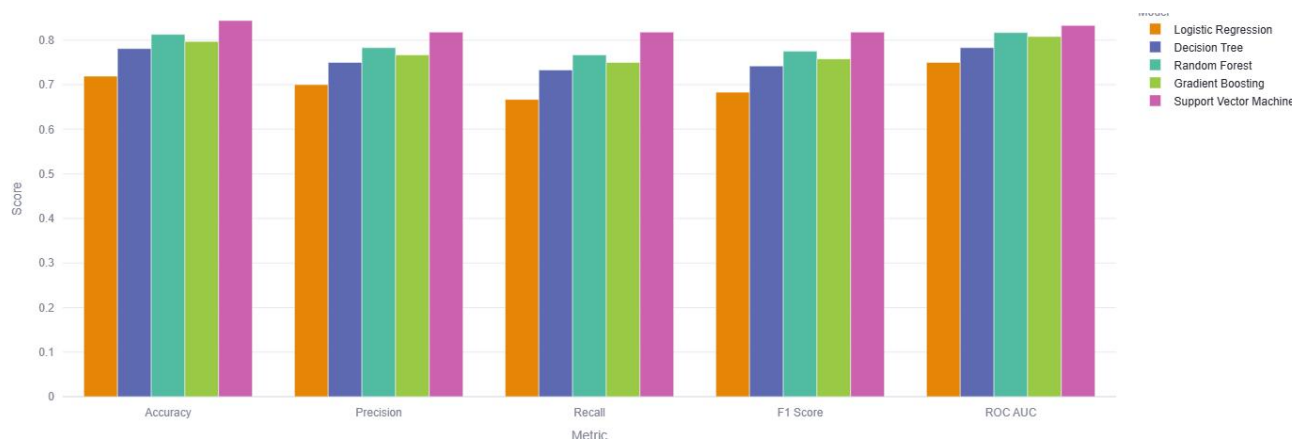


**Figure 2: The performance of the five classification algorithms compared.**

Figure 2 presents the performance of five machine learning models. Most of the metrics reported the highest performance of the SVM indicating its effectiveness in dealing with intricate patterns.

The comparative performance analysis of five machine learning algorithms, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting and Support Vector Machine (SVM), were carried out. Support Vector Machine (SVM) model with the highest overall performance with an accuracy of 84.4%, Precision of 81.8%, Recall of 81.8%, F1 Score of 81.8%, and ROC-AUC of 83.3%. Random Forest also performed well, with an accuracy of 81.3% and an ROC-AUC score of 81.7%, and balanced metric scores in all evaluation parameters. Decision Tree showed a moderate performance and Logistic Regression had the smallest scores when compared to other models. The outcome shows that the SVM model was the best model to identify the placement outcomes and to tackle the complexity of the dataset.

The comparative evaluation indicates that the Support Vector Machine achieved the highest overall predictive performance among the evaluated classifiers on the selected evaluation metrics. Random Forest and Gradient Boosting also demonstrated competitive performance, whereas Logistic Regression produced comparatively lower results. These findings suggest that non-linear machine learning models were better suited to capturing the relationships present in the dataset than simpler linear approaches. However, these results should be interpreted as predictive performance on the evaluated dataset rather than evidence of causal relationships between individual student characteristics and placement outcomes.

## 5.2 Confusion Matrix

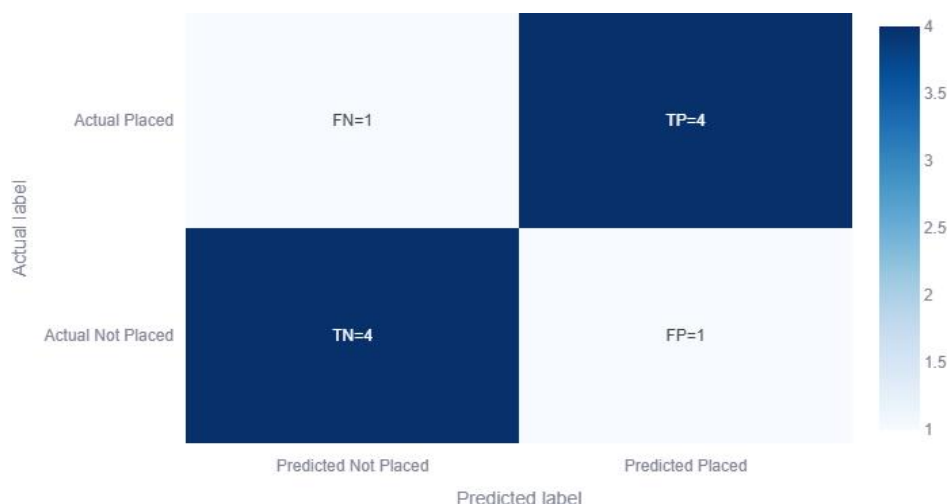
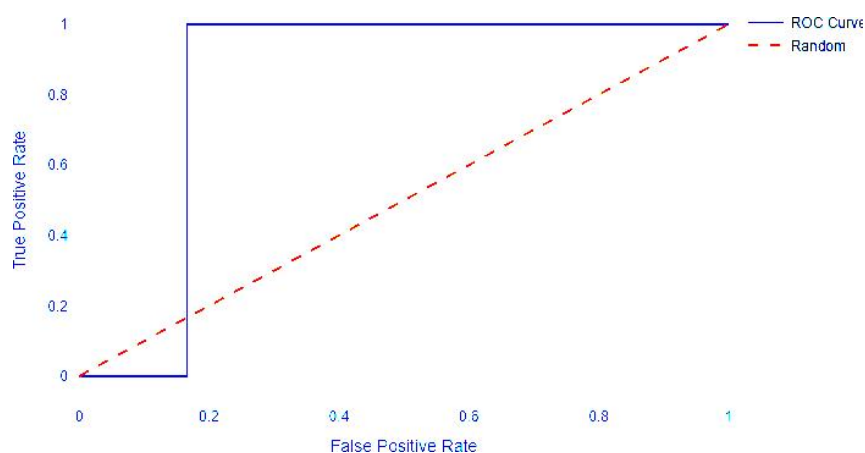


**Figure 3: Support Vector machine Classifier Confusion Matrix.**

As illustrated in figure 3, the SVM model assigned the correct classification of most cases belonging to either class meaning the balanced performance following the application of SMOTE.

### 5.3 ROC Curve



The SVM classifier has a ROC curve as shown in figure 4. It is positioned above the diagonal base which means that its classification performance is good with ROC-AUC = 0.833.

### 5.4 Feature Importance

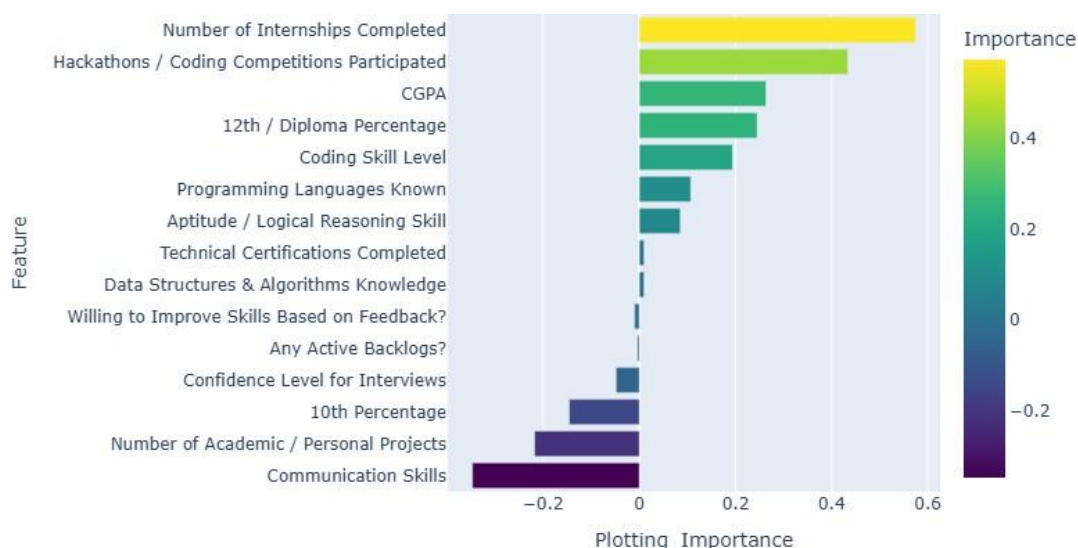


**Figure 5: Analysis Importance of Features, Depending on the magnitude coefficients of SVM.**

Figure 5 presents the relative importance of the input features used by the prediction framework. Since Support Vector Machines do not inherently provide conventional feature importance scores, the reported importance values should be interpreted as measures of feature contribution obtained from the model interpretation approach adopted in this study rather than direct evidence of causal influence. The results indicate that CGPA, technical skills, and internship experience were among the strongest predictors associated with placement outcomes within the analyzed dataset.

## Limitations

This study has several limitations. First, the predictive models were trained using a publicly available Kaggle dataset, which may not fully represent the characteristics of students from all educational institutions. Second, the institutional dataset collected from GNA University was used only to demonstrate the applicability of the trained prediction framework and was not incorporated into model training. Third, the study evaluated placement outcomes using the available predictor variables; additional factors such as socioeconomic background, extracurricular activities, aptitude test performance, and employer-specific recruitment criteria were not included. Future studies should validate the proposed framework using larger multi-institutional datasets to improve the generalizability of the results.

## 7. Ethical and Fairness Considerations

Student placement prediction systems should be used only as decision-support tools and should not replace human judgment during academic advising or recruitment processes. The proposed framework is intended to assist educational institutions in identifying students who may benefit from additional academic or career support rather than making automated placement decisions. Care should be taken to ensure that the model is periodically evaluated for potential bias across different student groups and updated using representative datasets to maintain fairness, transparency, and reliability. Conclusion and Scope for Future Enhancement

This study presented PlacePredict, a machine learning-based framework for predicting student placement outcomes using academic, technical, experiential, and soft-skill related attributes. Comparative evaluation of five supervised learning algorithms demonstrated that the Support Vector Machine achieved the best overall predictive performance on the evaluation dataset. The framework was subsequently applied to institutional student records to demonstrate its practical applicability. Although the results indicate that machine learning can effectively support placement prediction, additional validation using larger and more diverse institutional datasets is required before deployment in real-world educational environments. Future work will focus on expanding the dataset, incorporating additional student attributes, and improving model interpretability using explainable artificial intelligence techniques.

## References

- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning Analytics* (pp. 61–75). Springer.
- Bhoite, S., Deshmukh, M., & Patil, A. (2023). Probabilistic ensemble methods for student placement prediction. *International Conference on Educational Data Mining*.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

- Gupta, S., Verma, R., & Singh, A. (2025). Ensemble learning with feature importance analysis. *Expert Systems with Applications*, 242, 122847.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- International Journal of Artificial Intelligence in Education*, 35(1), 112–134.
- Journal of Computing in Higher Education*, 36(4), 512–531.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Kumar, M., Patel, S., & Gupta, R. (2023). Academic performance factors in student placement prediction. *International Journal of Educational Technology*, 15(2), 78–95.
- Romero, C., & Ventura, S. (2010). Educational data mining: A review. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 601–618.
- Tardalkar, S., Joshi, N., & Kulkarni, V. (2024). Random forest-based placement prediction system.
- Tiwari, S. K. (2025). Hybrid ensemble approach with explainable AI for student placement prediction.
- Varaprasad, K. V., Sharma, R., & Kumar, P. (2025). Comparative analysis of machine learning algorithms for campus placement prediction. *IEEE Transactions on Learning Technologies*, 18(3), 289–301.
- Yağci, M., Aydin, A., Dogan, B., & Bulut, E. (2022). Machine learning-based student placement prediction using academic performance and demographic features. *Journal of Educational Data Mining*, 14(2), 45–67.