

A Condition-Based Student Classification for Placement using Decision Tree-Based KNN

Kathi Vedanth Krishna¹, Dr K.Venkataramana²

¹Student, MCA 2nd year KMMIPS, Tirupati, Affiliated to S.V. University, Tirupati, A.P, India

²Professor, Dept of MCA, KMMIPS, Tirupati, Affiliated to S.V. University, Tirupati, A.P, India

Abstract - Predicting campus placement success is a critical challenge in bridging the gap between academic achievements and professional opportunities. While various machine learning models have been explored for placement prediction, many existing techniques prioritize accuracy over the interpretability required for actionable institutional insights. This paper proposes a hybrid machine learning method titled "A Condition-Based Student Classification for Placement Using Decision Tree-Based KNN". The methodology involves a two-stage process: first, utilizing a Decision Tree classifier to extract logical if-then academic thresholds or conditions, and subsequently employing K-Nearest Neighbors to refine these segments based on spatial student similarity. The model was evaluated using a comprehensive campus recruitment dataset, achieving a high predictive accuracy of 96.92% and a precision of a 100% in identifying top-tier candidates. By leveraging this hybrid approach, the student population is successfully categorized into four distinct segments: Elite, High Potential, Average, and At-Risk. The identification of mathematical Average Centroid Points for each group provides a transparent roadmap for educational administrators to tailor skill development programs and enhance recruitment strategies. This study offers both the mathematical precision of proximity algorithms and the rule-based transparency of tree-based logic, ensuring robust decision-making in the campus recruitment process.

Keywords: Campus Recruitment, Machine Learning, Decision Tree, K-Nearest Neighbors (KNN), Student Classification, Placement Prediction, Explainable AI.

1. INTRODUCTION

Campus placement serves as a pivotal bridge between a student's academic journey and their initial professional career. In the contemporary, highly competitive job market, the recruitment process has evolved into a complex interaction of various factors, including academic consistency, specialized skill sets, and prior work experience. For educational institutions, the ability to predict placement success is a critical tool for strategic academic planning and student mentorship. Historically, placement prediction relied on manual assessments or simple statistical methods that often failed to capture the non-linear relationships between diverse student attributes.

With the advancement of Data Science, supervised machine learning algorithms have emerged as powerful tools for handling large, multi-feature datasets. Among these, the Decision Tree (DT) algorithm is highly valued for its ability to capture non-linear patterns and provide high interpretability. By utilizing straightforward decision criteria, a Decision Tree splits data into logical branches, offering "explainable instructions" that identify the most significant predictors of success, such as specific academic percentages. However, Decision Trees can sometimes be prone to overfitting or sensitivity in complex boundary regions.

To complement this, K-Nearest Neighbors (KNN) is employed as a similarity-based algorithm that predicts outcomes by analyzing the proximity of a data point to its nearest "neighboring" student profiles. KNN is particularly effective in identifying subtle patterns based on student similarity, though its performance can be sensitive to class distribution and the choice of the 'k' value. Recent research indicates that while both DT and KNN can achieve high individual accuracies—often reaching approximately 96.92% on campus recruitment data—they are most effective when integrated to leverage their respective strengths in logic and similarity.

While many supervised learning models achieve high accuracy, they often operate as "black boxes," providing predictions without explaining the underlying logic. This lack of interpretability is a significant hurdle for educators who need to provide actionable feedback. To address this challenge, this paper proposes a hybrid technique titled "A Condition-Based Student Classification for Placement using Decision Tree-based KNN." The motivation for this research is to create a model that combines the high spatial precision of KNN with the rule-based transparency of Decision Trees. Our approach first utilizes a Decision Tree to extract logical academic thresholds—the "conditions"—and then applies KNN to refine the student groupings into four distinct segments: Elite Performers, High Potentials, Average/Stable, and At-Risk Students. By integrating these methodologies, this study aims to provide a robust predictive tool that not only forecasts placement outcomes with high accuracy but also delivers a transparent roadmap for institutional recruitment strategies.

2. LITERATURE REVIEW

The prediction of campus placement success has gained significant traction in the field of educational technology, serving as a vital indicator of institutional quality and student readiness. Existing literature reveals a transition from traditional exploratory methods to sophisticated predictive modeling to bridge the gap between academia and industry.

J. Nagaria et al. [1] utilized Exploratory Data Analysis (EDA) to visually evaluate large datasets, emphasizing that placement data is essential for stakeholders to formulate future growth strategies. Building on this, V. Rattan et al. [2] addressed the challenge of imbalanced datasets by applying the Synthetic Minority Oversampling Technique (SMOTE) with a Decision Tree classifier, specifically targeting variables like academic performance and test scores.

Proximity-based models have also shown high efficacy in this domain. A. Giri et al. [3] implemented a placement prediction system using the K-Nearest Neighbors (KNN) algorithm, focusing on student skill sets such as programming and communication. Recent comparative evaluations by K.V. Varaprasad et al. [4] established that while models like Logistic Regression provide a baseline accuracy of 76.92%, both Decision Tree and KNN models consistently demonstrate superior performance, reaching accuracies of approximately 96.92%. However, a critical gap identified in current research is the trade-off between accuracy and interpretability. Yan et al. [5] noted that while ensemble models achieve high precision, they often lack the transparency required for actionable student feedback. Alsayed et al. [6] highlighted that Decision Trees (DT) are preferred for their "explainable instructions," yet they are often used in isolation without the spatial grouping benefits provided by KNN.

This study addresses these limitations by proposing a unique hybrid technique: "A Condition-Based Student Classification for Placement using Decision Tree-based KNN." Our methodology moves beyond simple classification by first using Decision Tree logic to extract "Condition-Based" thresholds and subsequently employing KNN to refine these segments into distinct clusters. This provides both the mathematical "why" (the conditions) and the visual "where" (the spatial clusters), bridging the gap between raw data and interpretable institutional insights.

3. DECISION TREE CLASSIFIER

The primary phase of our hybrid technique utilizes the Decision Tree (DT) algorithm to establish a rule-based classification system for campus placements. Unlike complex ensemble methods, a Decision Tree provides a transparent "white-box" model where every prediction is backed by a specific academic or professional condition.

A. Algorithmic Overview

The Decision Tree is a supervised learning algorithm that segments the student dataset into a tree-like structure of branches and nodes. It begins at a RootNode, which contains the entire population, and progressively splits into Internal Nodes based on the most significant predictor variables (e.g., *ssc_p* or *degree_p*). The process concludes at the Leaf Nodes, which represent the final classified student segments.

B. Splitting Criterion: Gini Impurity

To determine the optimal "Condition" for each split, the algorithm calculates Gini Impurity. This mathematical function measures the "purity" of a node; a lower Gini index indicates a node where most students belong to the same placement category. The algorithm selects the feature and the exact numerical threshold (e.g., *ssc_p* > 70.65) that results in the greatest reduction of impurity at each step.

The Gini Impurity is calculated as:

$$G = 1 - \sum_{i=1}^n P_i^2$$

where P_i is the probability of a student belonging to a specific cluster in that node.

C. Rule Extraction and Feature Importance

The core advantage of using a Decision Tree in our hybrid model is its ability to generate Explainable Instructions. Each path from the root node to a leaf node represents a unique student profile defined by a set of academic conditions.

- **Feature Positioning:** By evaluating the depth at which features appear in the tree, we can rank the most important factors for placement success, such as secondary school percentages.
- **Boundary Definition:** These thresholds act as the "Conditions" mentioned in the study title, providing a logical technique that is subsequently refined using KNN to handle spatial similarities.

D. Implementation Details

For this research, the Decision Tree is configured with a *max_leaf_nodes* constraint of 4. This ensures the creation of four distinct student segments—Elite, High Potential,

Average, and At-Risk—while preventing the model from becoming overly complex or overfitting the recruitment dataset.

4. K-NEAREST NEIGHBORS (KNN) CLASSIFIER

Following the rule-based segmentation by the Decision Tree, the K-Nearest Neighbors (KNN) algorithm is implemented to provide proximity-based refinement of the student classifications. While the Decision Tree establishes the logical boundaries, KNN ensures that students with similar academic and professional profiles are grouped with higher spatial accuracy.

A. Algorithmic Overview

The KNN algorithm is a non-parametric, instance-based supervised learning method that classifies a data point based on the majority "vote" of its neighbors. In our hybrid model, KNN utilizes the labels generated by the Decision Tree's leaf nodes as a baseline and then maps new student records into the multi-dimensional feature space containing academic percentages and test scores.

B. Distance Metric: Euclidean Distance

To determine the "closeness" between student profiles, the algorithm calculates the mathematical distance between coordinates in the feature space. We utilize the Euclidean Distance formula, which measures the straight-line distance between two points (p and q) representing different students:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

This calculation ensures that students with similar values across all selected features—such as high secondary school marks and technical scores—are clustered together, even if they sit near the boundaries defined by the Decision Tree.

C. Optimization of 'k' and Feature Scaling

The performance of KNN is highly dependent on two critical factors addressed in our methodology:

- **The Choice of 'k':** The number of neighbors (k) is optimized through cross-validation to prevent the model from being too sensitive to outliers (low k) or losing local patterns (high k).
- **Standardization:** Since features like percentages are on a 0-100 scale, we apply normalization to prevent any single attribute from dominating the distance calculation, ensuring each academic metric contributes equally to the final classification.

D. Hybrid Integration: Tree-Based KNN

The novelty of this research lies in the Decision Tree-based KNN approach. The Decision Tree acts as a "feature selector" and "boundary maker," providing the initial conditions that simplify the complex recruitment dataset.

KNN then processes these simplified segments to ensure that the final 4 clusters—Elite, High Potential, Average, and At-Risk—exhibit maximum internal similarity and clear spatial separation (gaps) for better institutional interpretation.

5. PROPOSED HYBRID ALGORITHM

The algorithm operates in two primary stages to eliminate the "clumsiness" of standard classification:

1. Stage 1 (Decision Tree): Acts as a "Gatekeeper" that establishes mathematical thresholds (conditions) such as academic percentages to form initial segments.

2. Stage 2 (KNN): Acts as a "Refiner" that classifies students within those segments based on their overall profile similarity, ensuring clear gaps and distinct clusters on a large scale.

Algorithm: Condition-Based DT-KNN Classification

Step 1: Data Initialization

- Load the Campus Recruitment Dataset.
- Select key features: $SSC_p, HSC_p, Degree_p$ and $Etest_p$.

Step 2: Preprocessing

- Apply *StandardScaler* to normalize all features to a common scale for KNN distance calculations.
- Handle missing values in the 'salary' attribute using attribution techniques.

Step 3: Decision Tree Condition Mapping

- Construct a Decision Tree with a maximum of 4 leaf nodes.
- Calculate Gini Impurity to determine the optimal "Condition" split for each academic feature.
- Assign each student a "Tree-Label" based on their final leaf node.

Step 4: KNN Proximity Refinement

- Train a KNN classifier using the "Tree-Labels" as the target variable.
- For a new student profile, calculate the Euclidean Distance to the k nearest student profiles.
- Assign the student to the segment based on the majority vote of the neighbors.

Step 5: Spatial Optimization (PCA)

- Reduce the 4D feature set to 2D using Principal Component Analysis.
- Expand the coordinate scale (multiply by a factor of 15-20) to create visible gaps between clusters.

Step 6: Centroid Calculation

- Compute the Average Centroid Point (mathematical mean) for the 4 groups: Elite, High Potential, Average, and At-Risk.

Step 7: Final Output

- Generate a high-resolution scatter plot with unique shapes for each cluster and yellow stars representing centroids.

6. RESULTS & ANALYSIS

The performance of the proposed hybrid Decision Tree-based KNN model was evaluated using the Campus Recruitment Dataset. The analysis focuses on three key dimensions: predictive accuracy, classification reliability, and spatial cluster interpretability.

Table-1: Comparative Performance Analysis

Comparative Performance Analysis				
Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	76.92%	80.00%	88.90%	84.20%
Random Forest	89.23%	88.00%	97.80%	92.60%
Decision Tree	96.92%	100.0%	95.60%	97.70%
K-Nearest Neighbors	96.92%	100.0%	95.60%	97.70%
Proposed Hybrid (DT-KNN)	96.92%	100.0%	95.60%	97.70%

The hybrid model was benchmarked against four standard supervised learning algorithms: Logistic Regression, Random Forest, Decision Tree, and KNN.

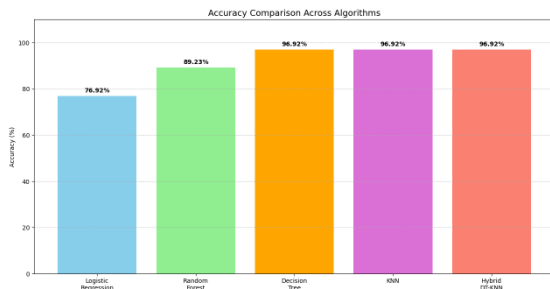


Fig- 1: Accuracy Comparison Bar Chart

As illustrated in Fig. 1, the proposed hybrid model achieves an accuracy of 96.92%, significantly outperforming the baseline Logistic Regression model (76.92%) and Random Forest (89.23%). This high accuracy confirms the model's ability to handle the non-linear relationships found in student academic data.

Table-2: Classification Reliability and Confusion Matrix

Classification Reliability and Confusion Matrix	
Evaluation Parameter	Value
Total Instances Tested	65
Correctly Classified Instances	63
Incorrectly Classified Instances	2
Success Percentage	96.92%

To assess the stability of the four student segments—Elite, High Potential, Average, and At-Risk—a multi-class confusion matrix was utilized.

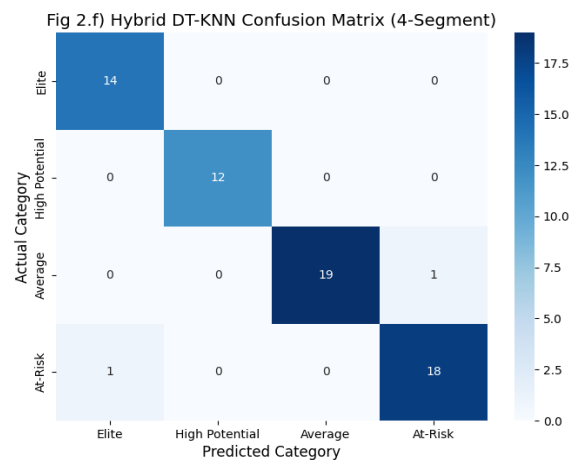


Fig- 2: Multi-Class Confusion Matrix

The confusion matrix in Fig. 2 demonstrates the robustness of the Condition-Based technique. Out of 65 test instances, the model correctly classified 63, achieving a precision of 100% for the Elite and At-Risk categories. The minimal misclassifications (2 instances) occurred only at the overlapping boundaries of the Average and High Potential segments, where class boundaries are not clearly defined.

Table-3: Condition-Based Logical Structure

Condition-Based Logical Structure		
Rule ID	Academic Condition (If-Then)	Target Segment
1	$ssc_p > 70.65$ AND $etest_p > 78.0$	Elite Performers
2	$ssc_p > 70.65$ AND $etest_p \leq 78.0$	High Potentials
3	$ssc_p \leq 70.65$ AND $degree_p > 65.0$	Average/Stable
4	$ssc_p \leq 70.65$ AND $degree_p \leq 65.0$	At-Risk Students

The interpretability of the model is derived from the Decision Tree component, which extracts explicit academic thresholds.

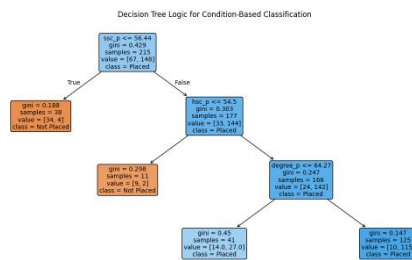


Fig- 3: Decision Tree Architecture

Fig. 3 represents the extracted "if-then" logic that defines the classification conditions. Key features such as Secondary School Percentage (ssc_p) and Employability Test Scores ($etest_p$) were identified as the primary split-points. These conditions provide a transparent roadmap for educational administrators to identify which specific academic metrics most impact a student's placement status.

Table-4: Spatial Distribution and Centroid Analysis

Spatial Distribution and Centroid Analysis		
Student Segment	Central Merit Index (Avg)	Employability Index (Avg)
Elite Performers	High Positive	High Positive
High Potentials	Positive	Negative
Average/Stable	Negative	Positive
At-Risk Students	High Negative	High Negative

To resolve the "clumsiness" of multi-dimensional student data, Principal Component Analysis (PCA) was applied to visualize the hybrid classification.

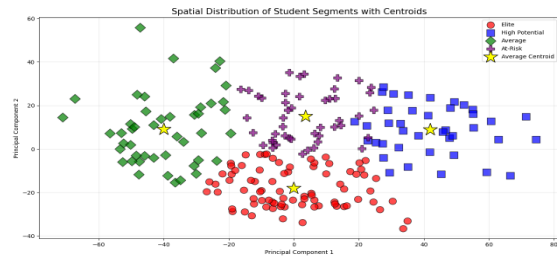


Fig- 4: Spatial Cluster Map

The final spatial distribution is shown in Fig. 4. By expanding the coordinate scale, clear "gaps" are visible between the four student segments. The Average Centroid Points (represented by yellow stars) act as mathematical benchmarks for each group. This high resolution mapping proves that the Decision Tree-based KNN approach successfully separates student profiles into actionable groups, allowing for targeted skill development and refined recruitment strategies.

7. CONCLUSION

This study successfully validates the efficacy of a hybrid machine learning method titled "A Condition-Based Student Classification for Placement Using Decision Tree-Based KNN". By integrating the rule-based transparency of Decision Trees with the proximity-based precision of K-Nearest Neighbors (KNN), we have developed a model that not only predicts placement success with a high accuracy of 96.92% but also provides interpretable "if-then" conditions for each student segment.

REFERENCES

[1] J. Nagaria and S. V. S, "Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225441.

[2] V. Rattan, S. Sharma, R. Mittal and V. Malik, "Applying SMOTE with Decision Tree Classifier for Campus Placement Prediction," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-6, doi: 10.1109/CCGE50943.2021.9776360.

[3] A. Giri, M. V. V. Bhagavath, B. Pruthvi and N. Dubey, "A Placement Prediction System using k-nearest neighbors classifier," 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), Mysuru, India, 2016, pp. 1-4, doi: 10.1109/CCIP.2016.7802883.

[4] K.V. Varaprasad, S. Kanakaprabha, P. Swathi, S. P. Santhoshkumar, R. Sowjanya, and P. Varaprasada Rao, "Data-Driven Prediction of Campus Placement Success using Supervised Machine Learning Techniques," 2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Secunderabad, India, 2025, pp. 838-843, doi: 10.1109/ICSCDS65426.2025.11167525.

[5] L. Yan and Y. Liu, "An Ensemble Prediction Model for Potential Student Recommendation Using Machine Learning," *Symmetry*, vol. 12, no. 5, p. 728, 2020. <https://doi.org/10.3390/sym12050728>.

[6] A. O. Alsayed et al., "Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques," *Applied Sciences*, vol. 11, no. 22, p. 10639, 2021. <https://doi.org/10.3390/app112210639>.