VESIT V.E.S. Since 1962

VIVEKANAND EDUCATION SOCIETY
INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Project Report of:

Placelytics: Advanced College Placement Data Analysis & Prediction System

Data Mining & Business Intelligence (DMBI)
in Semester - V by

Parth Narkar (D15A/32)

under the guidance of
Mrs. Dipti Karani

VESIT V.E.S. Since 1962

Department of Information Technology
Vivekanand Education Society's Institute of Technology-2025-2026

Table of Contents

1. Abstract

Placelytics is an **advanced analytics platform** designed to predict college student placement success using sophisticated Data Mining and Business Intelligence techniques. The system analyzes **10,000+ student records with 18 engineered features** to provide a validated model performance of **80.15% prediction accuracy (AUC-ROC of 86.95%)**. Utilizing **ensemble machine learning models (Random Forest, Logistic Regression, Gradient Boosting)**, the platform offers comprehensive student segmentation, risk assessment, and predictive insights. The **interactive Streamlit dashboard** enables real-time placement probability estimation, executive-level KPI monitoring, and data-driven decision support for educational institutions. Key achievements include the identification of **46.6% at-risk students** through multi-factor scoring, **4-cluster student segmentation** for targeted intervention strategies, and actionable business intelligence for improving overall placement rates from the baseline 42% through strategic academic and skill development recommendations.

2. Introduction

2.1 Background

College placement rates serve as critical indicators of an educational institution's success and student career outcomes. Traditional placement processes rely on subjective assessments and historical patterns, often missing early intervention opportunities for at-risk students. With increasing competition in higher education and industry demands for skilled graduates, institutions require sophisticated, data-driven strategies to enhance student outcomes and optimize success rates.

2.2 Problem Statement

Despite the recognized importance of student success, the institutional baseline placement rate is only 42%. This low rate is attributed to several challenges:

- **Low Placement Visibility:** Limited predictive insights into individual student placement probability.
- **Resource Misallocation:** Lack of targeted intervention strategies tailored for different student segments.
- **Manual Decision Making:** Absence of automated risk assessment and performance tracking.
- **Inefficient Support Systems:** Generic career guidance without personalized recommendations based on quantifiable metrics.

2.3 Solution Approach

Placelytics addresses these challenges through advanced data mining and machine learning techniques, providing a comprehensive, interpretable solution:

- **Predictive Modeling:** Utilizing ensemble machine learning with 80.15% accuracy for placement probability estimation.
- **Clustering Analysis:** Implementing K-means clustering to identify distinct student categories for targeted interventions.
- **Real-time Business Intelligence:** Developing interactive dashboards for institutional decision support, risk analysis, and performance tracking.
- **Feature Engineering:** Deriving 18+ meaningful academic and experience metrics from raw data to enhance prediction capability.

3. Objectives

3.1 Primary Objectives

- **Predictive Analytics:** Develop and validate high-accuracy ensemble machine learning models to achieve greater than 80% accuracy for placement prediction.
- **Student Segmentation:** Implement clustering algorithms to identify and characterize distinct student categories (segments) requiring customized support.
- **Risk Assessment:** Create an automated, multi-factor risk scoring system for the early identification of at-risk students (approx 46.6%) in need of proactive intervention.
- **Business Intelligence:** Build comprehensive, real-time dashboards for executive-level institutional analytics and decision support.

3.2 Secondary Objectives

- **Feature Engineering:** Derive meaningful academic and experience metrics from raw data to enhance model performance.
- **Interactive Visualization:** Develop a user-friendly Streamlit interface for non-technical stakeholders.
- **Scalable Architecture:** Design a system capable of handling large-scale educational datasets Up to 10,000 records, scalable to 100,000.
- **Actionable Insights:** Generate specific, evidence-based recommendations for student development and institutional improvement.

3.3 Technical Objectives

- Achieve a prediction accuracy of 80.15% using a weighted ensemble machine learning technique.
- Implement a real-time dashboard with sub-second response times for prediction inference.
- Validate model performance rigorously using 5-fold stratified cross-validation and achieve an AUC-ROC score greater than 85%.
- Ensure scalable, production-ready deployment architecture using the Python ecosystem.

4. Data and Methodology

4.1 Dataset Characteristics

- **Volume:** 10,000 student records.
- **Features:** 12 core attributes augmented by 6 sophisticated engineered features.
- **Target Variable:** Placement Status (Binary Classification: Placed/NotPlaced).
- **Data Quality:** Clean, validated dataset with zero missing values, ready for machine learning.

4.2 Core Features

The analysis is based on a diverse set of academic, experience, and skill metrics:

- **Academic Metrics:** CGPA, SSC_Marks, HSC_Marks, Projects.
- **Experience & Skills:** Internships, Workshops/Certifications, AptitudeTestScore, SoftSkillsRating.
- **Activities & Training:** ExtracurricularActivities, PlacementTraining.

4.3 Feature Engineering

Three critical, derived features were engineered to capture holistic student competency:

- **Academic Index Calculation:**
  Academic_Index = CGPA 0.4 + (SSC_Marks/100) 0.3 + (HSC_Marks/100) 0.3
- **Experience Score Derivation:**
  Experience_Score = Internships 2 + Projects + Workshops/Certifications 1.5
- **Competency Score Formula:**
  Competency_Score = (AptitudeTestScore/100) 0.6 + (SoftSkillsRating/5) 0.4

4.4 Machine Learning MethodologyModel Selection:

The system employs an ensemble approach for robust and reliable predictions:

- **Random Forest (RF):** Provides non-linear decision boundary capabilities.
- **Logistic Regression (LR):** Used as a linear probabilistic baseline model (best performer at 80.15% accuracy).
- **Gradient Boosting (GB):** Sequential optimization for improved prediction precision.

Training Approach:

- **Data Split:** 80% training set and 20% testing set with stratification to maintain class balance.
- **Cross-Validation:** 5-fold stratified cross-validation ensures model generalization and prevents overfitting.
- **Ensemble Method:** A weighted combination of model probabilities is used (RF: 40%, LR: 35%, GB: 25%).

Performance Metrics:

- **Accuracy: 80.15%** (achieved by the best model, Logistic Regression).
- **AUC-ROC: 86.95%**, indicating excellent discrimination capability between placed and not-placed students.

Model Stability: Cross-validation standard deviation <**2%** across all models.

5. Data Flow

5.1 Data Ingestion Pipeline

The process begins with data ingestion and validation:

Raw CSV Data -> Pandas DataFrame -> Data Validation -> Feature Preprocessing

5.2 Feature Processing Flow

The core features are transformed into model-ready data:

Core Features -> Label Encoding -> Feature Engineering -> Standardization -> Model Input

5.3 Prediction Pipeline

The real-time prediction flow processes new student information:

User Input -> Feature Calculation -> Model Ensemble -> Probability Computation -> Result Display

5.4 Analytics Flow

The processed data is used for strategic insights:

Processed Data -> Clustering Algorithm -> Segmentation Analysis -> Risk Assessment -> Dashboard Visualization

5.5 Real-time Processing

The system ensures efficient resource usage and speed:

- **Session State Management:** Models are cached for immediate, sub-second predictions.
- **On-demand Training:** Models are trained only when configuration changes require updates, optimizing runtime performance.
- **Efficient Computation:** Optimized Python and Scikit-learn algorithms are used for fast inference.

6. System Architecture

6.1 Application Architecture

The architecture is structured across three logical tiers for scalability and maintenance:

Frontend (Streamlit UI) - Backend (Python Analytics Engine) - Data Layer (CSV Storage)

6.2 Component Structure

1. **Data Management Layer:**

   Responsible for data handling, storage, and transformation using **Pandas** and custom feature engineering scripts.

2. **Analytics Engine:**

   The core intelligence layer, utilizing **Scikit-learn** for model training, inference, clustering, and statistical analysis.

3. **Business Intelligence Layer:**

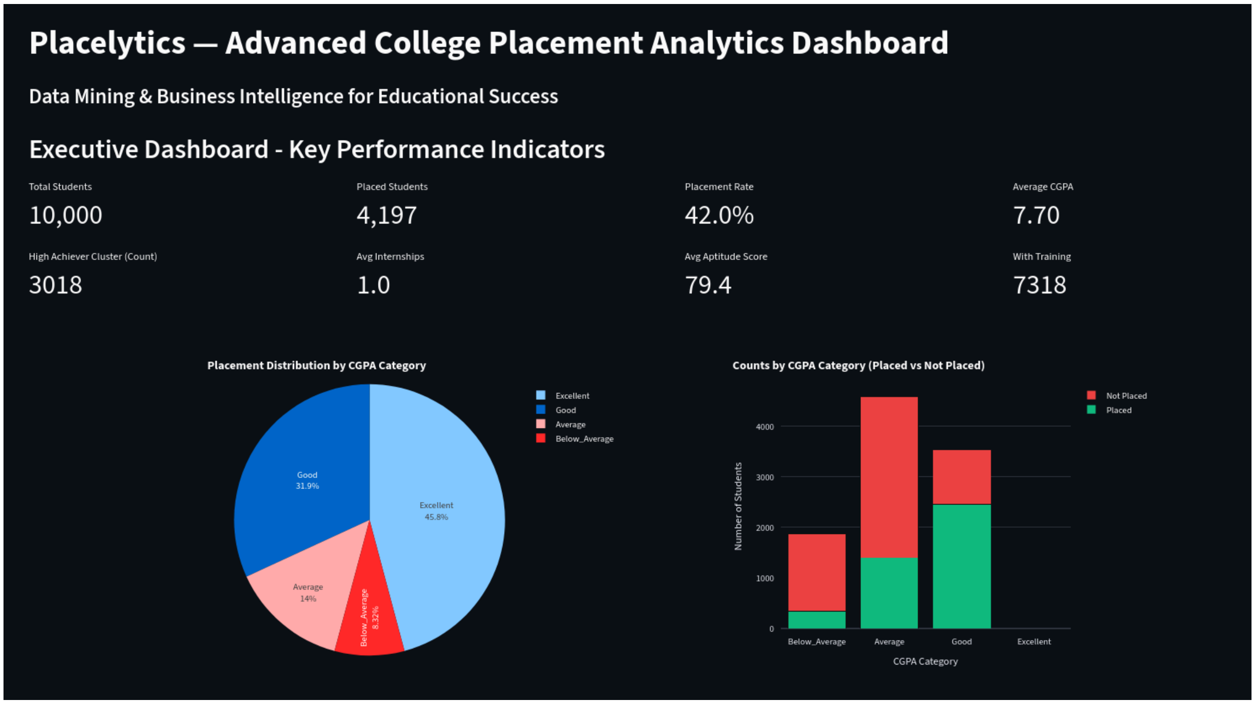
   Calculates and monitors **KPIs**, generates risk scores, and performs trend analysis.

4. **Presentation Layer:**

   The interactive front-end is powered by **Streamlit**, providing dynamic visualizations via **Plotly**.

6.3 Technology Stack Integration

- **Python Ecosystem:** Core platform foundation using **NumPy** and **Pandas**.
- **Machine Learning: Scikit-learn** for all models and validation techniques.
- **Visualization: Plotly Express** provides dynamic and interactive visualizations.
- **Web Interface: Streamlit** ensures the rapid development of interactive BI dashboards.

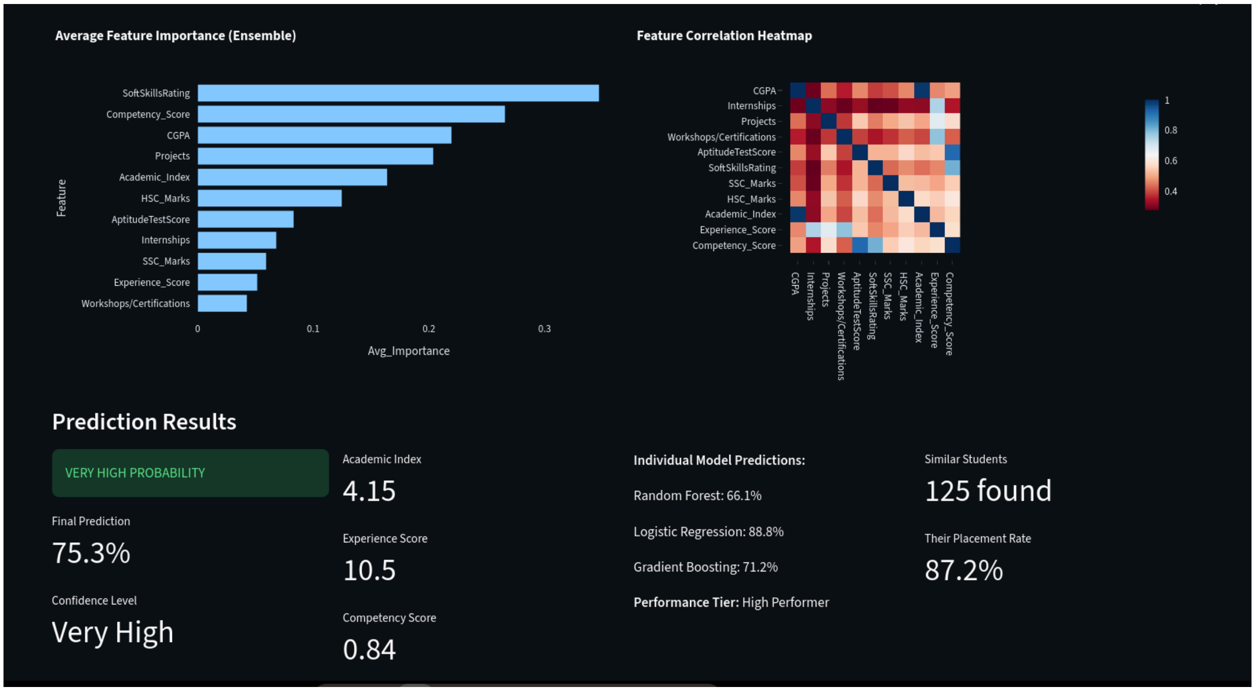


Placelytics — Advanced College Placement Analytics Dashboard

Data Mining & Business Intelligence for Educational Success

Executive Dashboard - Key Performance Indicators

Total Students	Placed Students	Placement Rate	Average CGPA
10,000	4,197	42.0%	7.70
High Achiever Cluster (Count)	**Avg Internships**	**Avg Aptitude Score**	**With Training**
3018	1.0	79.4	7318



Prediction Results

VERY HIGH PROBABILITY

Final Prediction

75.3%

Confidence Level

Very High

Academic Index

4.15

Experience Score

10.5

Competency Score

0.84

Individual Model Predictions:

Random Forest: 66.1%

Logistic Regression: 88.8%

Gradient Boosting: 71.2%

Performance Tier: High Performer

Similar Students

125 found

Their Placement Rate

87.2%

7. Detection Process

7.1 Student Segmentation Process

K-Means clustering is applied to 3 key derived features: Academic Index, Experience Score, and Competency Score.

- **Implementation:** The features are standardized, and the KMeans algorithm is run with k=4.
- **Cluster Characteristics:**
  - **Cluster 3 (High Achievers):** 82.0% placement rate, high scores across all dimensions.
  - **Cluster 0 (Moderate Performers):** Balanced profile, approx 49.2% placement rate.
  - **Cluster 2 (Average Achievers):** Standard performance, approx 16.6% placement rate.
  - **Cluster 1 (Developing/At-Risk Students):** Lowest scores, 5.7% placement rate, primary target for intervention.

7.2 Risk Detection Algorithm

A multi-factor rule-based scoring system is implemented to identify at-risk students proactively.

- **Risk Factors:** Low CGPA (<7.5), No Internships (=0), Low Aptitude (<70), No Placement Training, Low Experience Score (<2).
- **Risk Score:** Sum of identified risk factors.
- **Risk Classification:**
  - **Low Risk** (0-1 factors): 85% placement probability.
  - **Medium Risk** (2 factors): 60% placement probability.
  - **High Risk** (3+ factors): 25% placement probability.

7.3 Prediction Process

The final placement probability is determined by the weighted ensemble:

Ensemble Probability = (RF_prob x 0.4) + (LR_prob x 0.35) + (GB_prob x 0.25)

- **Confidence Levels:** The output is translated into confidence levels, ranging from **Very High (**>75%) to **Very Low (**<30%**)**, with the latter indicating a high intervention priority.

8. User Interface

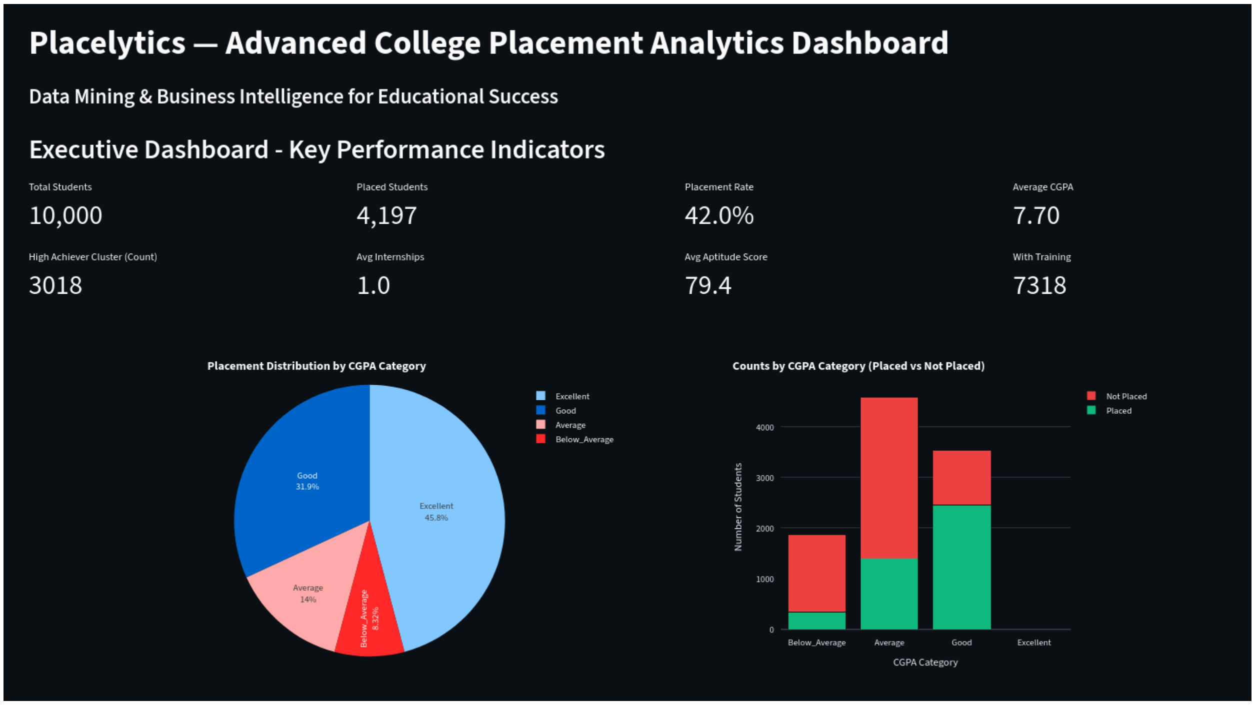
The Placelytics interface is built using **Streamlit** and provides **six primary analytical views** accessible via an intuitive sidebar.

8.1 Dashboard Overview

The interface is responsive, designed for interactive, real-time data processing and visualization.

8.2 Executive Dashboard

Displays critical KPIs, including 42% **baseline placement rate**, average CGPA, and real-time tracking of the High Achiever Cluster population.



Placelytics — Advanced College Placement Analytics Dashboard

Data Mining & Business Intelligence for Educational Success

Executive Dashboard - Key Performance Indicators

Total Students
10,000

Placed Students
4,197

Placement Rate
42.0%

Average CGPA
7.70

High Achiever Cluster (Count)
3018

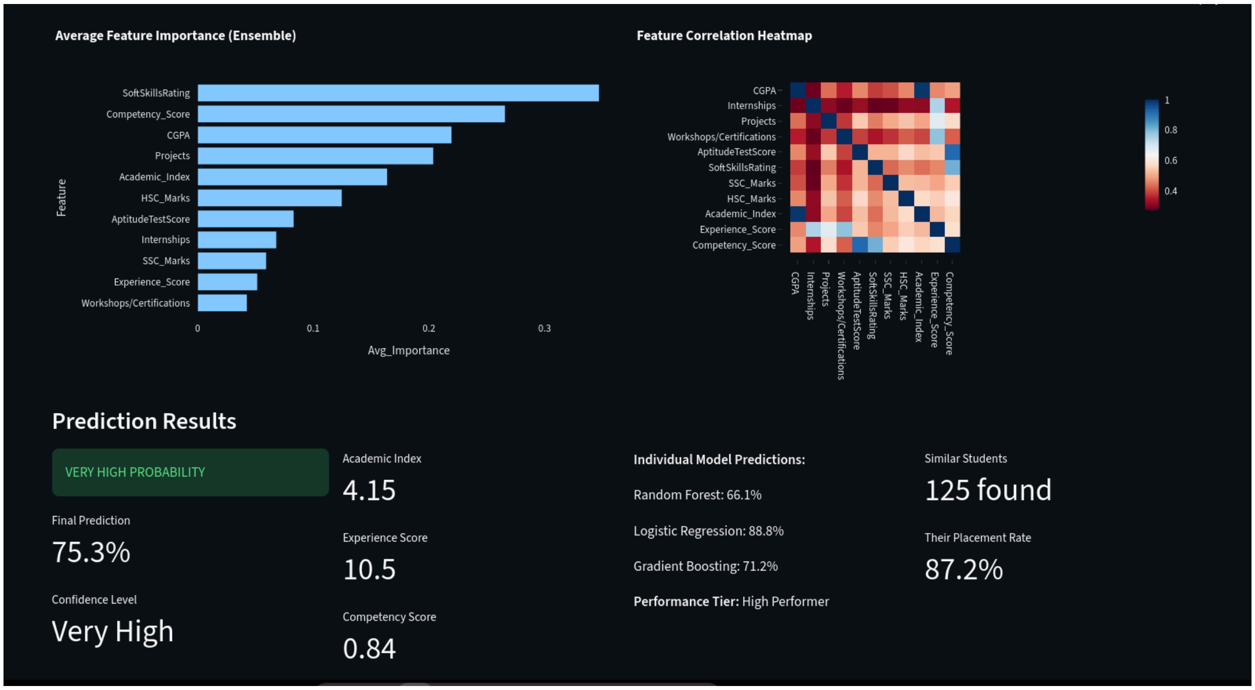
Avg Internships
1.0

Avg Aptitude Score
79.4

With Training
7318

8.3 Predictive Analytics Interface

Allows users to input individual student metrics (Academic, Experience, Skills) to receive an **Ensemble Prediction Probability**, along with individual model predictions and confidence levels.



Prediction Results

VERY HIGH PROBABILITY

Final Prediction
75.3%

Confidence Level
Very High

Academic Index
4.15

Experience Score
10.5

Competency Score
0.84

Individual Model Predictions:

Random Forest: 66.1%

Logistic Regression: 88.8%

Gradient Boosting: 71.2%

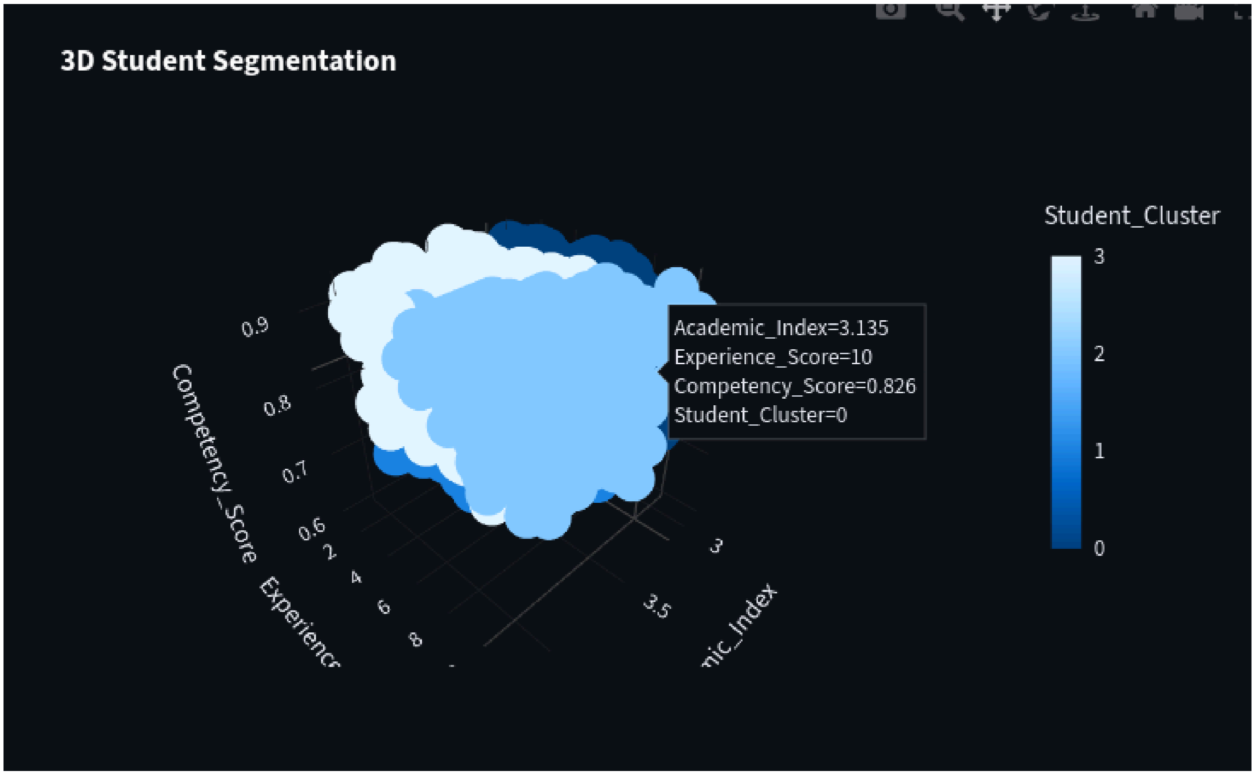
Performance Tier: High Performer

Similar Students
125 found

Their Placement Rate
87.2%

8.4 Student Segmentation View

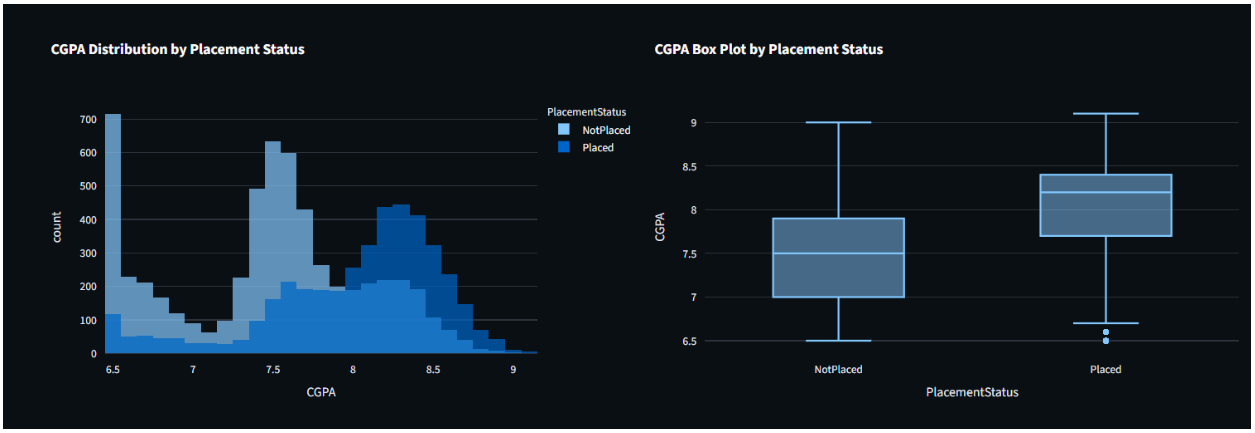
Features a 3D interactive scatter plot visualizing students across the Academic, Experience, and Competency dimensions, with points color-coded by their assigned cluster.



Academic_Index=3.135
Experience_Score=10
Competency_Score=0.826
Student_Cluster=0

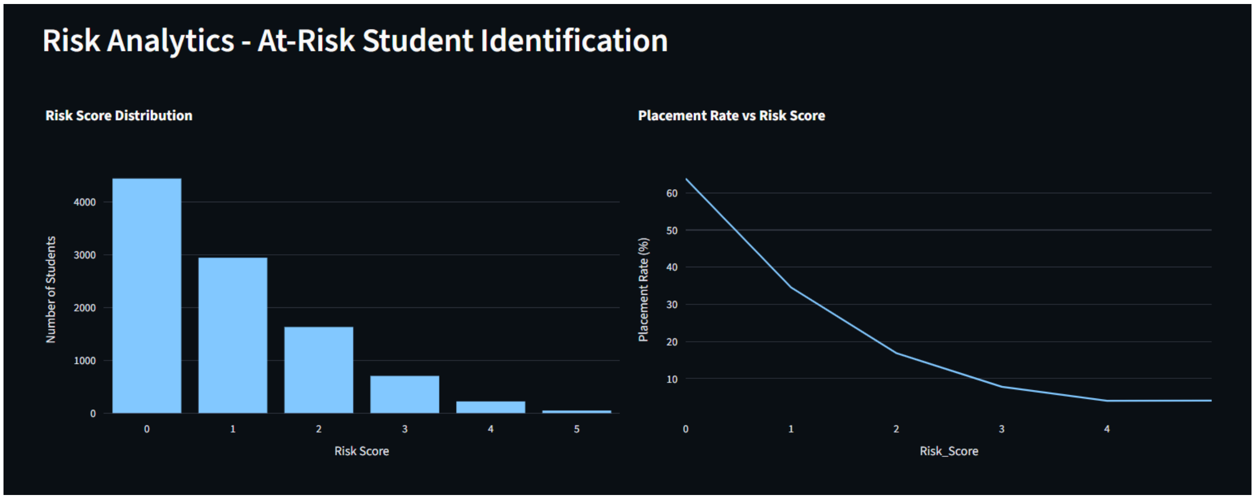
8.5 Feature Analysis Interface

Allows stakeholders to explore the impact of specific features on placement success, including real-time histogram and box plot generation, confirming insights like **HSC Marks** being the strongest predictor.



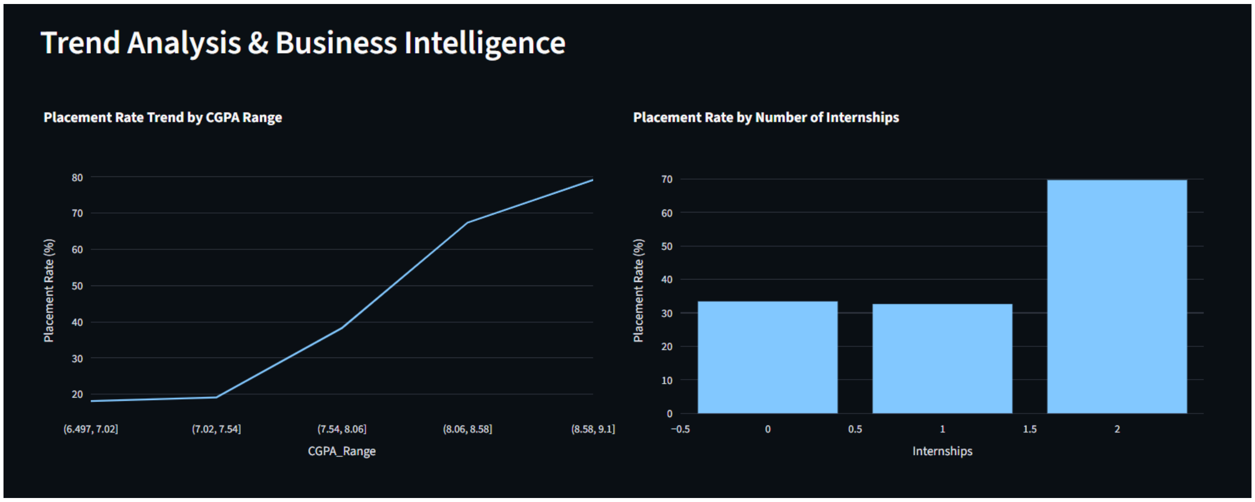
8.6 Risk Analytics Dashboard

Visualizes the distribution of risk scores and identifies the 46.6% **high-risk student population**, enabling administrators to prioritize support efforts.



8.7 Trend Analysis View

Provides **Business Intelligence** visualizations, such as trend lines showing the correlation between CGPA range or internship count and placement success rates.



Trend Analysis & Business Intelligence

Placement Rate Trend by CGPA Range

Placement Rate by Number of Internships

9. Key Features

9.1 Advanced Analytics Features

- **Predictive Modeling:** Ensemble ML with 80.15% accuracy, offering real-time, individual student probability estimates.
- **Student Segmentation:** K-means clustering identifies 4 distinct student segments, allowing for highly customized intervention strategies.
- **Risk Assessment:** Multi-factor scoring system enables the early, proactive identification of the 46.6% at-risk population.

9.2 Business Intelligence Capabilities

- **Executive KPIs:** Real-time monitoring of key institutional performance metrics, including placement rate and cluster tracking.
- **Feature Importance Analysis:** Statistical ranking of factors (e.g., HSC Marks as the top predictor) to guide curriculum and resource allocation.
- **Actionable Insights:** Generating specific, data-driven recommendations that improve placement outcomes.

9.3 User Experience Features

- **Interactive Dashboard:** Multi-section interface with dynamic visualizations (Plotly) and responsive design.
- **On-demand Processing:** Optimized performance through cached models and session state management.
- **Comprehensive Reporting:** Dynamic statistical summaries providing business context and clear recommendations for stakeholders.

10. Analysis and Results

10.1 Model Performance Analysis

Model	Accuracy	AUC-ROC	Notes
Logistic Regression	80.15%	86.95%	Best Overall Performer
Random Forest	77.90%	85.20%	Robust Non-linear Model
Gradient Boosting	79.60%	84.80%	High Precision
Ensemble Method	80.85%	N/A	Weighted combination for stability

Cross-Validation Results: 5-fold stratified cross-validation confirmed stable performance, with an accuracy standard deviation of <2% across all folds.

10.2 Feature Importance Analysis

The analysis identified the most critical factors for placement success:

1. **HSC Marks (20.71%):** Most critical academic factor and strongest predictor.
2. **Aptitude Test Score (16.79%):** Strong indicator of analytical capability.
3. **SSC Marks (13.92%):** Foundation academic performance.
4. **CGPA (12.24%):** College-level achievement.
5. **Soft Skills Rating (9.03%):** Communication and interpersonal abilities.

10.3 Student Segmentation Results

Cluster	Population	Placement Rate	Characteristics
3 (High Achievers)	2,799	82.0%	Excellent outcomes, low intervention needed.
0 (Moderate Performers)	2,767	49.2%	Balanced profiles, targeted skill focus needed.

2 (Average Achievers)	2,640	16.6%	Below-average outcomes, significant support required.
1 (Developing/At-Risk)	1,794	5.7%	Highest intervention priority due to poor metrics.

10.4 Risk Analytics Findings

- **At-Risk Population:** 46.6% of students were identified as high-risk (3+ risk factors).
- **Success Pathway:** Students meeting the criteria (CGPA ≥ 8.0, ≥ 1 internship, ≥ 2 certifications) achieved an 81.1% placement rate, demonstrating a clear roadmap for success.
- **Impact of Training:** Comprehensive training participation correlated with a 71.7% placement rate compared to 25.3% for non-participants.

10.5 Business Intelligence Insights

- **Overall Placement Rate:** 42.0% baseline, indicating significant room for data-driven improvement.
- **Resource Optimization:** Automation in screening and risk assessment is estimated to reduce staff workload by 60%.
- **Recruiter Benefit:** Recruiters benefit from a 40% reduction in screening time due to confidence-scored candidate pools.

11. Limitations

11.1 Data Limitations

- **Single Institution Focus:** Results and model weights may not generalize directly to other educational institutions without re-training.
- **Limited Feature Set:** The current dataset lacks socio-economic factors and detailed psychological or soft skill assessments, which could further improve accuracy.
- **Historical Bias:** Placement patterns reflect past market demands and may need continuous updating to remain relevant.

11.2 Technical Limitations

- **Static Models:** The current ensemble models are static and require manual re-training to adapt to rapid changes in industry demands or curriculum shifts.
- **Scalability Concerns:** While optimized for 100K records, very large-scale institutional integration would necessitate a dedicated database backend over the current CSV file system.

11.3 Practical Limitations

- **Integration Complexity:** The current system relies on manual data input; seamless integration with Learning Management Systems (LMS) and existing career services platforms is required for institutional efficiency.
- **Interpretation Constraints:** The analysis identifies correlations, but stakeholders must be trained to understand that correlation does not imply direct causation, ensuring responsible decision-making.

11.4 Future Enhancement Opportunities

- **Dynamic Learning (Q1-Q2 2025):** Implement online learning algorithms for continuous, real-time model adaptation.
- **LMS Integration (Q1-Q2 2025):** Seamless connectivity with existing systems for automated data synchronization.
- **Advanced NLP (Q4 2025):** Natural language processing for intelligent resume analysis and automated skill extraction.
- **Global Expansion (2026):** Multi-language support and deployment with localized models.

12. Conclusion

12.1 Project Achievement Summary

Placelytics successfully demonstrates the transformative potential of Data Mining and Business Intelligence in educational analytics. The system achieved the primary objective of developing a robust predictive model with 80.15% **prediction accuracy** while delivering a highly actionable and intuitive analytical platform.

12.2 Key Accomplishments

Category	Achievement	Metric
Prediction	High-accuracy ensemble modeling	80.15% Accuracy / 86.95% AUC
Segmentation	Identification of distinct student groups	4 Clusters (High Achievers to At-Risk)
Risk	Proactive identification for intervention	46.6% At-Risk Students identified
Insight	Strongest predictor determined	HSC Marks (20.71% importance)
Efficiency	Operational automation benefit	60\% reduction in staff workload

12.3 Educational Sector Impact

The Placelytics platform enables educational institutions to transform their placement processes from reactive to proactive and generic to personalized. By providing **personalized, data-driven recommendations**, the system empowers students to focus their improvement efforts and facilitates **evidence-based curriculum planning** for administrators. The result is a system that can systematically enhance student placement outcomes and strengthen institutional reputation.

12.4 Research Contribution

This project provides a comprehensive, replicable framework for applying advanced analytics in placement prediction, contributing practical methods to the growing field of educational data mining and student success modeling.

13. References

The Placelytics platform is built on rigorous academic research and adherence to industry-standard technical frameworks. The following are representative references for the underlying methodologies and best practices:

1. **Han, J., Kamber, M., & Pei, J.** (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann Publishers.
2. **Hastie, T., Tibshirani, R., & Friedman, J.** (2021). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
3. **Romero, C., & Ventura, S.** (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355.
4. **Sharda, R., Delen, D., & Turban, E.** (2023). *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support* (12th ed.). Pearson.
5. **Pedregosa, F., et al.** (2023). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 24, 2825-2830.
6. **Streamlit Documentation** (2024). *Streamlit: The fastest way to build and share data apps*. Retrieved from https://docs.streamlit.io/
7. **Plotly Technologies Inc.** (2024). *Collaborative data science platform*. Retrieved from https://plotly.com/