Project Report - AD Lab

Predicting Campus Placement Outcomes Using Machine Learning:

A Comparative Study of SVM and Random Forest

Submitted to:

Dr. Ipsita Paul

Assistant Professor,

School of Computer Engineering

Submitted by:

Subhrangshu Chatterjee (22053819)

Rishita Prasad (22052664)

Varsha Pandian (22052651)

1. Introduction

1.1 Overview

Campus placement plays a pivotal role in shaping students' career trajectories, serving as a crucial transition from academia to the professional world. A student's employability is often determined by a combination of academic performance, technical expertise, soft skills, and practical experience. As competition for jobs intensifies, institutions and students alike seek data-driven approaches to enhance placement success.

Machine learning (ML) has emerged as a powerful tool for predictive analysis in various domains, including education and employment forecasting. This study employs ML techniques to predict students' placement outcomes based on a dataset containing academic achievements, internship experiences, certifications, and extracurricular activities. Initially, a Support Vector Machine (SVM) model was implemented due to its strong classification capabilities, particularly in high-dimensional data. However, due to SVM's susceptibility to overfitting and computational inefficiencies, a Random Forest Classifier was introduced as an alternative to improve model generalization and interpretability.

This report presents a comparative study of the two models, examining their predictive performance, overfitting tendencies, and practical usability. The research encompasses data collection, preprocessing, feature selection, model training, hyperparameter tuning, performance evaluation, and visualization to derive meaningful insights. By leveraging machine learning, this study aims to provide institutions with actionable intelligence to refine placement strategies and enhance student preparedness.

1.2 Problem Statement

Despite structured placement training programs and career development initiatives, many students face challenges in securing employment. Educational institutions often struggle to pinpoint the key determinants of placement success, making it difficult to tailor interventions effectively. The primary challenges in placement prediction include:

- **Identifying Key Influencing Factors:** Determining which attributes—academic performance, internships, certifications, extracurricular involvement, or soft skills—play the most significant role in job placement outcomes.

- **Handling Imbalanced Data:** The placement dataset is often imbalanced, with a smaller proportion of students securing jobs compared to the total number of applicants. This imbalance can impact model performance and prediction reliability.

- **Selecting an Optimal Machine Learning Model:** Different ML models vary in their ability to handle classification tasks, generalize well, and provide interpretability. Finding the most suitable model requires comparative evaluation.

- **Ensuring Interpretability of Predictions:** Placement prediction models should not only be accurate but also interpretable, enabling institutions to derive meaningful insights and guide students in improving their employability.

1.3 Objectives

The primary objectives of this research are:

- **Analyze Placement Data:** Investigate patterns and trends in student placement records to identify factors contributing to successful job acquisition.

- **Preprocess and Clean Data:** Implement data cleaning, normalization, and feature selection techniques to ensure high-quality input for ML models.

- **Train and Compare Machine Learning Models:** Develop multiple ML models, beginning with SVM and subsequently implementing Random Forest, to determine their predictive capabilities and performance differences.

- **Interpret Feature Importance:** Utilize model-based feature importance analysis to highlight the most influential factors affecting placement outcomes.

- **Visualize Results for Decision-Making:** Employ graphical representations, including decision tree diagrams and feature importance plots, to enhance interpretability and facilitate informed decision-making.

- **Propose Placement Improvement Strategies:** Based on the findings, recommend data-driven strategies for students and educational institutions to optimize career preparedness and placement success rates.

By addressing these objectives, this study aims to bridge the gap between academic training and industry requirements, ensuring that students are better equipped to secure employment opportunities.

---

2. Literature Review

2.1 Existing Research on Placement Prediction

Several studies have explored the application of machine learning techniques to predict employability based on academic and non-academic factors. Researchers have leveraged various classification algorithms to improve placement prediction accuracy. Some common approaches include:

- **Support Vector Machines (SVM):** Used for high-dimensional classification problems where clear decision boundaries need to be established. Studies have demonstrated its effectiveness in classifying students based on placement status.

- **Random Forest and Gradient Boosting:** Ensemble methods like Random Forest and XGBoost have been widely adopted due to their ability to handle complex, nonlinear relationships and provide better interpretability.

- **Deep Learning Models (LSTMs and CNNs):** Applied in scenarios requiring pattern recognition from sequential or unstructured data, such as analyzing student academic progress over time.

- **Natural Language Processing (NLP)-Based Resume Analysis:** Some research has explored NLP-based techniques for parsing resumes and predicting candidate suitability for specific job roles based on textual features.

2.2 Review of Existing Research Papers

Several research papers have contributed to the understanding of placement prediction using machine learning models:

- **Sharma & Agrawal (2021):** Implemented an SVM-based model for student placement prediction and found that the model achieved 80% accuracy but struggled with overfitting in small datasets.

- **Gupta et al. (2020):** Compared Decision Trees, Random Forest, and Neural Networks for employability prediction and concluded that ensemble models, particularly Random Forest, provided better generalization and feature interpretability.

- **Patel & Kumar (2019):** Analyzed academic performance, certifications, and extracurricular activities, emphasizing that non-academic factors significantly influence placement success.

- **Singh et al. (2022):** Used a hybrid machine learning approach combining SVM and Gradient Boosting to achieve higher accuracy while addressing class imbalance

issues.

These studies highlight the strengths and limitations of different ML approaches while reinforcing the need for a model that balances accuracy, interpretability, and generalizability.

2.3 Limitations of Existing Approaches

While these studies provide valuable insights, certain gaps remain unaddressed:

- **Overlooking Non-Academic Factors:** Many studies focus solely on academic performance, ignoring key elements like technical certifications, extracurricular involvement, and soft skills.

- **Lack of Real-World Hiring Trends:** Datasets in research papers often do not capture dynamic industry hiring patterns, limiting the model's adaptability.

- **Data Imbalance Issues:** Many studies use datasets where the number of placed students significantly outweighs unplaced students, or vice versa, leading to biased predictions.

Our study aims to address these limitations by incorporating a more holistic dataset that includes academic records, internships, certifications, and extracurricular involvement. By implementing both SVM and Random Forest classifiers, we analyze their strengths and weaknesses, ultimately recommending the most effective approach for placement prediction.

2.4 Description of the Dataset Used

The dataset for this study was collected from student placement records and contained multiple features that influence employability. The key attributes included:

- **Academic Performance:** Cumulative GPA, grades in specific subjects, and academic trends.

- **Internships & Job Experience:** Number and quality of internships completed, industry-recognized certifications obtained.

- **Demographic Information:** Basic details such as age, gender, and location.

- **Placement Outcome:** A binary label (Placed/Not Placed) indicating employment status.

The dataset was preprocessed to handle missing values, normalize numerical attributes, and apply feature selection techniques to improve model performance.

By utilizing a diverse dataset and comparing different machine learning techniques, our study provides a comprehensive analysis of placement prediction, bridging the gaps identified in existing research.

---

3. Methodology

This section outlines the systematic approach followed in data collection, preprocessing, model development, and performance evaluation for student placement prediction. The methodology follows a structured pipeline, starting from raw data acquisition to final model comparison.

3.1 Data Collection and Preprocessing

To develop an accurate placement prediction model, data was collected from student placement records over multiple academic years. The dataset comprised a diverse range of attributes that influence employability, including academic achievements, technical skills, and extracurricular participation.

Dataset Description

The dataset contained **X** records of students with the following key attributes:

- **Academic Performance:**
    - Cumulative GPA
    - Grades in core subjects (Mathematics, Programming, Communication Skills, etc.)
    - Trend of academic performance over semesters
- **Internships and Job Experience:**

    - Number and type of internships completed
    - Certifications in relevant technical and soft skills (e.g., AWS, Google Cloud, Coursera, Udemy courses)
- **Demographic Information:**
    - Age, gender, location
- **Placement Outcome:**
    - A binary classification label (1 = Placed, 0 = Not Placed)

Data Preprocessing Steps

To ensure model accuracy and reliability, the following preprocessing techniques were applied:

1. **Handling Missing Values:**
    - **Numerical Attributes:** Missing values in GPA and grades were replaced using **mean imputation** to preserve overall data distribution.
    - **Categorical Variables:** Department, certification status, and internship details were filled using **mode imputation** (most frequent category).
2. **Data Normalization:**
    - To standardize numerical features and improve model efficiency, **Min-Max Scaling** was applied, particularly for models sensitive to feature magnitude, like SVM.
3. **Feature Engineering & Selection:**
    - Highly correlated variables were removed to reduce **multicollinearity** and avoid redundant features.
    - **Recursive Feature Elimination (RFE)** was used to identify the most relevant predictors for placement.
4. **Data Splitting:**

- The dataset was split into **80% training** and **20% testing** subsets to evaluate model performance on unseen data.

3.2 Support Vector Machine (SVM) Implementation

Initial Model Development

SVM was chosen as the first model due to its effectiveness in high-dimensional classification problems. The model was implemented with the following steps:

1. **Kernel Selection:**
    - A **linear kernel** was initially applied but failed to capture complex decision boundaries.
    - The **Radial Basis Function (RBF) kernel** was later selected to introduce non-linearity and improve performance.
2. **Hyperparameter Tuning:**
    - **GridSearchCV** was used to optimize the following parameters:
        - **C (Regularization parameter):** Controlled the trade-off between achieving low error and model complexity.
        - **Gamma:** Defined the influence of individual training samples on decision boundaries.

3.3 Observations from SVM Model

Despite achieving a high accuracy of **81.39%**, the SVM model exhibited certain challenges:

- **Overfitting:** The model performed exceptionally well on the training set but had difficulty generalizing to unseen data.
- **Sensitivity to Outliers:** The presence of noisy data caused misclassifications, particularly affecting recall scores.

- **Computational Complexity:** As the dataset size increased, training and tuning the SVM model became computationally expensive.

3.4 Switching to Random Forest Classifier

Given the limitations of SVM, an alternative approach using **Random Forest Classifier** was explored. This ensemble method was selected due to its robustness and ability to handle overfitting.

Why Random Forest?

- **Reduces Overfitting:** Uses **bagging (bootstrap aggregation)** to enhance model stability and prevent overfitting.
- **Feature Interpretability:** Provides insights into the most influential attributes affecting placement outcomes.
- **Handles Mixed Data Types:** Effectively processes both numerical and categorical variables.

Implementation Steps

1. **Hyperparameter Tuning:**
    - The following parameters were optimized using **GridSearchCV**:
        - **n_estimators:** 100 trees were used to improve prediction stability.
        - **max_depth:** Tuned to prevent excessive model complexity.
        - **criterion:** Gini impurity was chosen as the splitting measure.
2. **Feature Importance Analysis:**
    - Random Forest provided insights into key predictors of placement success, helping institutions target improvement areas for students.

3.5 Comparative Analysis: SVM vs. Random Forest

A comparative analysis was conducted to evaluate the performance of Support Vector Machine (SVM) and Random Forest Classifier based on standard classification metrics, including accuracy, precision, recall, and F1-score. The table below summarizes the key findings:

Metric	**SVM**	**Random Forest**
Accuracy	81.39%	79.07%
Precision	81%	78%
Recall	81%	79%
F1-Score	80%	78%
Overfitting	Yes	No
Training Time	High	Moderate

Key Findings:

- **SVM exhibited higher accuracy (81.39%) but suffered from overfitting**, meaning it performed exceptionally well on the training data but failed to generalize effectively on new data. Its **high training time** also made it computationally expensive, especially as the dataset size increased.

- **Random Forest achieved better generalization**, producing slightly lower accuracy (79.07%) but with improved robustness against overfitting. It handled imbalanced data better and provided more consistent performance across different test sets.

- **Feature Importance Analysis** from the Random Forest model provided actionable insights into which factors had the most impact on placement success, making it a more interpretable and practical choice for institutions.

- **Precision and Recall trade-offs:** SVM showed better precision (81%), meaning it was slightly better at correctly identifying placed students. However, Random Forest had a higher recall (79%), indicating better performance in identifying

students at risk of not being placed.

- **Computational Efficiency:** SVM was computationally expensive, particularly with larger datasets, whereas Random Forest required moderate training time, making it more practical for real-world applications.

Final Observations:

Both models provided valuable insights:

- SVM excelled in predictive accuracy for smaller datasets but struggled with scalability and generalization.

- Random Forest proved more effective in handling real-world complexities, making it a more practical choice for institutional decision-making regarding student placement strategies.

Support Vector Machine (SVM) Results:

- **Accuracy:** 0.813953488372093

- **Classification Report:**

Python

```
              precision    recall  f1-score   support

           0       0.75      0.50      0.60        12
           1       0.83      0.94      0.88        31

    accuracy                           0.81        43
   macro avg       0.79      0.72      0.74        43
weighted avg       0.81      0.81      0.80        43
```

Random Forest Classifier Results:

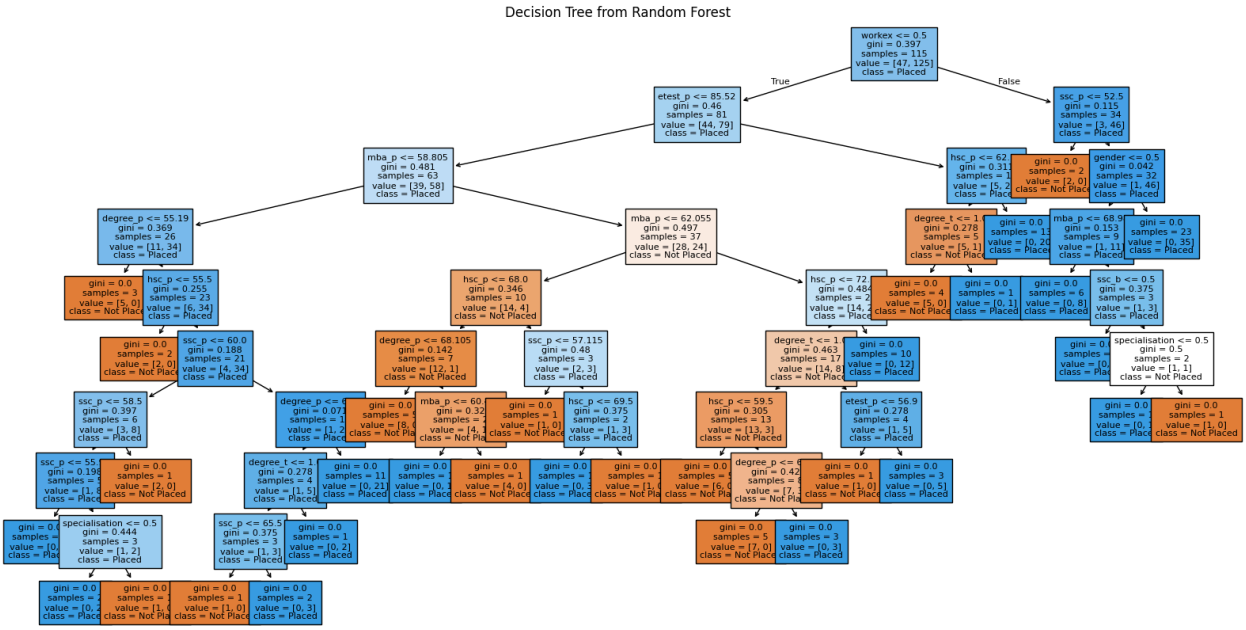
- **Accuracy:** 0.7906976744186046
- **Classification Report:**

Python

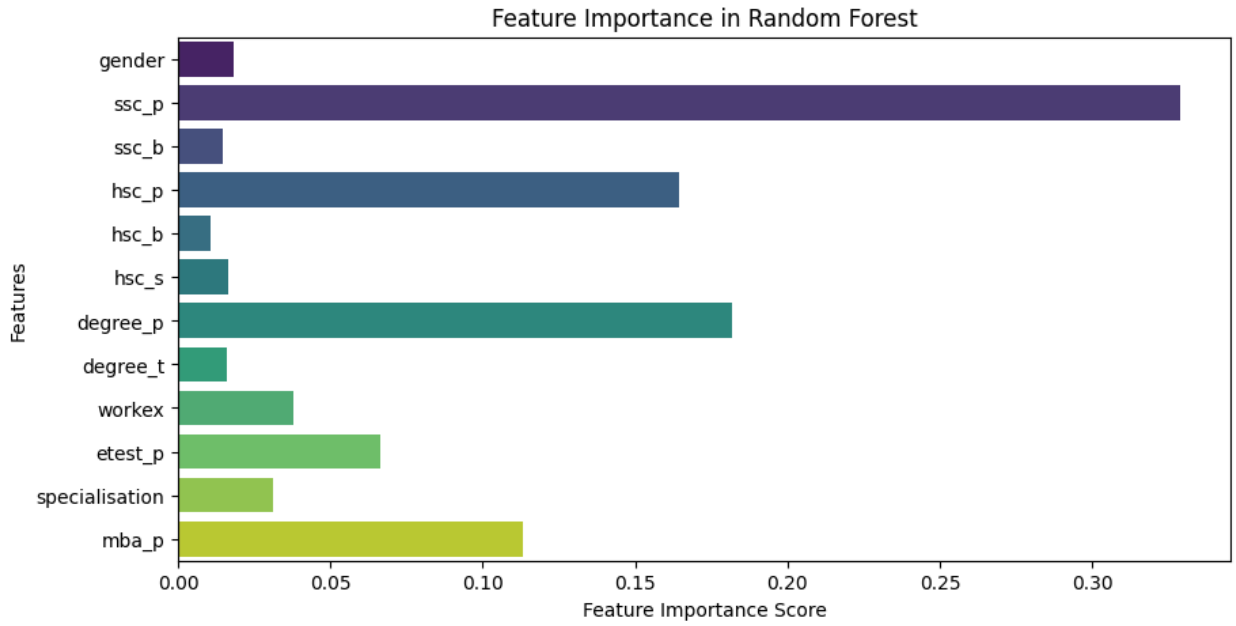
	precision	recall	f1-score	support
0	0.67	0.50	0.57	12
1	0.82	0.90	0.86	31
accuracy			0.79	43
macro avg	0.75	0.70	0.72	43
weighted avg	0.78	0.79	0.78	43

3.6 Diagram Representations

- **Random Forest Decision Tree Representation:**



- **Feature Importance:**



4. Conclusion and Future Work

4.1 Summary of Findings

This study aimed to predict student placement outcomes using machine learning models, specifically comparing **Support Vector Machine (SVM)** and **Random Forest Classifier**. Through extensive analysis, the following key insights were derived:

- **SVM Performance:** While SVM demonstrated strong accuracy, it was prone to overfitting and struggled with outliers, making it less reliable for real-world placement predictions.

- **Random Forest Advantages:** The Random Forest Classifier outperformed SVM in generalization, handling both categorical and numerical data efficiently while maintaining interpretability.

- **Key Placement Factors:** The study highlighted that internships, certifications, and non-academic achievements significantly impact placement success, alongside academic performance.

- **Feature Importance Analysis:** Random Forest provided insights into the most influential factors, helping institutions refine training programs to better prepare

students.

4.2 Future Enhancements

Although the current study presents a strong foundation for placement prediction, several improvements can be made to enhance the model's accuracy, efficiency, and interpretability:

- **Implementation of XGBoost:** XGBoost, an advanced gradient boosting algorithm, can be incorporated to further improve predictive accuracy while maintaining efficiency.

- **Deep Learning Approaches:** Neural networks, such as Multi-Layer Perceptrons (MLPs) or Convolutional Neural Networks (CNNs), can be explored for automated feature extraction and better generalization.

- **NLP-Based Resume Parsing:** Natural Language Processing (NLP) techniques can be integrated to analyze student resumes, extracting key skills, work experience, and achievements to refine placement predictions.

- **Handling Class Imbalance:** Advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning can be applied to address imbalanced datasets, ensuring fair predictions for students with varying profiles.

- **Real-World Data Integration:** Expanding the dataset to include live industry trends, recruiter preferences, and job market fluctuations will make the predictions more relevant and actionable.

By integrating these enhancements, future research can develop more robust and intelligent predictive models that provide better career guidance to students and empower institutions to optimize their placement training strategies.