Student Placement Prediction Report

Name: Japjeet kaur

Id: C0937720

1. Introduction

This project aims to predict student placement status using academic and demographic features from a dataset containing student profiles and placement outcomes. Various classification models were trained and evaluated to identify the best-performing algorithm for predicting whether a student will be placed or not.

2. Dataset Overview

- Dataset: Placement_Data_Full_Class.csv
- Number of instances: 215
- Features: Academic percentages (SSC, HSC, Degree, MBA, etc.), demographic info (gender, work experience), and placement status.
- Target variable: `status` — indicates whether a student was placed (1) or not (0).

Initial Data Inspection

- No missing values found.
- Categorical variables encoded using LabelEncoder.
- Dropped irrelevant columns: `sl_no`, `salary`.

3. Exploratory Data Analysis (EDA)

Distribution of Placement Status

The dataset shows a moderate balance between placed and not placed students.

Key Feature Insights

- Academic scores (`ssc_p`, `hsc_p`, `degree_p`, `mba_p`) generally show higher averages for placed students.
- Boxplots reveal that degree and MBA percentages are strong indicators of placement.
- Work experience positively influences placement likelihood.

Correlation Analysis

- Positive correlations observed between `mba_p`, `degree_p`, `ssc_p` and placement status.
- Heatmap and pairplots visually confirmed these relationships.

4. Model Building and Evaluation

Multiple classification models were trained on 70% of the data and tested on the remaining 30%. Performance was evaluated using accuracy, precision, recall, and F1-score.

Models Tested

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Gradient Boosting Classifier
- Decision Tree Classifier
- Voting Classifier (Ensemble of multiple models)

Baseline Results

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	78.5%	79.2%	93.3%	85.7%
Random Forest	83.1%	82.7%	95.6%	88.7%
SVM	84.6%	81.8%	100%	90.0%
KNN	83.1%	81.5%	97.8%	88.9%
Gradient Boosting	83.1%	84.0%	93.3%	88.4%
Decision Tree	73.8%	78.0%	86.7%	82.1%
Voting Classifier	84.6%	84.3%	95.6%	89.6%

- SVM achieved the highest recall (perfectly identified all placed students) and the best F1 score.
- The Voting Classifier provided the most balanced performance, combining strengths of individual models.

5. Hyperparameter Tuning

GridSearchCV was used to optimize key models for better performance, focusing on maximizing F1-score.

Best Parameters Found

Model	Best Parameters
Random Forest	{'n_estimators': 50, 'max_depth': None, 'min_samples_split': 5, 'min_samples_leaf': 1}
SVM	{'C': 1, 'kernel': 'linear', 'gamma': 'scale'}
Gradient Boosting	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50}

Results After Tuning

Model	Accuracy	Precision	Recall	F1 Score
Random Forest (Tuned)	86.2%	87.5%	93.3%	90.3%
SVM (Tuned)	80.0%	82.0%	91.1%	86.3%
Gradient Boosting (Tuned)	83.1%	84.0%	93.3%	88.4%
Voting Classifier (with tuned models)	83.1%	82.7%	95.6%	88.7%

- Tuned Random Forest improved in all metrics and became the best performing model overall.
- Surprisingly, tuning SVM reduced its accuracy slightly, possibly due to parameter constraints.
- Voting Classifier maintained strong performance with tuned base models.

6. Conclusions

- Academic performance, especially degree and MBA percentages, are strong predictors of student placement.
- Work experience also positively correlates with placement likelihood.
- Among all models, Random Forest (tuned) achieved the best balance of precision, recall, and overall accuracy.
- The Voting Classifier provides a robust ensemble approach combining multiple models for balanced results.
- The high recall values across models indicate reliable identification of placed students, crucial for real-world placement prediction.