

XGBoost-Based Career Domain Prediction Using Academic Performance and Socioeconomic Features: An Empirical Study

Kirti Khare

Department of Computer Science, Kalinga University, Naya Raipur, Chhattisgarh, India

Dr. Anupa Sinha

Department of Computer Science & IT, Kalinga University, Naya Raipur, Chhattisgarh, India

Abstract

Accurate career prediction for students can significantly improve educational outcomes and reduce graduate underemployment. This paper presents an empirical study comparing five machine learning classifiers – Naive Bayes, Decision Tree, Logistic Regression, Random Forest, and XGBoost – for predicting student career domains across six categories: Technology, Healthcare, Business & Finance, Arts & Humanities, Law & Administration, and Engineering. A dataset of 750 student records was constructed using a structured survey collecting academic performance scores, technical and interpersonal skill ratings, and demographic attributes from undergraduate and postgraduate students in Chhattisgarh, India. Feature preprocessing included one-hot encoding, min-max normalization, and SMOTE-based class balancing within 10-fold stratified cross-validation. XGBoost achieved the highest accuracy of 81.3% with F1-score of 0.807 and AUC-ROC of 0.924. Feature importance analysis identified CGPA, Technical Skills Score, and Interpersonal Skills Index as the top three predictors. These findings confirm gradient boosting ensemble methods as the optimal approach for multi-class career prediction in heterogeneous Indian student populations.

Keywords

XGBoost, Career Prediction, Academic Performance, Machine Learning, SMOTE, Indian Higher Education, Feature Importance, Random Forest, Educational Data Mining.