



Career Mapping and Enhancing Personalized Education through Machine Learning-Based Recommendation Systems

Angeline R¹, Charanya R², Konduru Sri Abhinaya² and Lingutla Shasank Chowdary²

¹ Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Tamil Nadu, India

² Department of Computer Science and Engineering with specialization in Artificial Intelligence and Machine Learning, SRM Institute of Science and Technology, Ramapuram, Tamil Nadu, India
angelinr1@srmist.edu.in

Abstract. A machine learning-based predictive model was developed to deliver personalized career recommendations based on student data, including academic performance and personal attributes. The dataset includes information such as gender, absenteeism, extracurricular activities, part-time work status, study habits, and subject scores. Machine learning techniques such as Logistic Regression, Support Vector Classifier, Random Forest, K-Nearest Neighbors, CatBoost, LightGBM, and XGBoost — were trained and tested to determine the most effective approach for career prediction. To improve model performance, Bayesian optimization was used to optimize algorithm parameters. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized to ensure a balanced representation of career aspirations. The system accepts user inputs, processes them through the trained model, and returns the top five career recommendations with associated probability. Feature scaling was implemented to normalize input data, further improving prediction accuracy. This study evaluates the use of machine learning in career guidance to assist students in making informed decisions. Future additions could incorporate dynamic features and adapt the model to evolving educational landscapes, further improving its usefulness in career counseling.

Keywords: Machine learning, Recommendation system, SMOTE, CatBoost, Classification, Bayesian Optimization

1 Introduction

Choosing a career path is one of the most significant decisions a student will make, but the overwhelming amount of alternatives and lack of personalized assistance can make the process challenging. Traditional career counseling generally takes a one-size-fits-all approach, which may fail to address each student's

unique talents and interests. To bridge this gap, this study introduces a machine learning-driven career recommendation system that analyzes a student's academic performance, personal traits, and extracurricular involvement to suggest suitable career paths. By offering personalized recommendations, the system helps students gain clearer insights into their options and make well-informed decisions. Additionally, to improve accuracy, the model tackles challenges like class imbalance using techniques such as Synthetic Minority Over-sampling (SMOTE) and fine-tunes its performance through hyperparameter optimization.

2 Related Work

Many researchers have explored the use of machine learning in recommendation systems for career guidance and personalized education. Guntupalli et al. [1] introduced Career Compass, a system that applies machine learning to help individuals choose career paths based on their unique attributes and industry trends. Similarly, Siswipraptini et al. [2] developed a career recommendation model tailored specifically for IT students in Indonesia, demonstrating its effectiveness in improving career decision-making. In the realm of personalized learning, Bin et al. [3] conducted an extensive study on recommendation systems in e-learning platforms, emphasizing various machine learning techniques that enhance student engagement. Kamal et al. [4] provided a systematic review of academic recommender systems, discussing their role in influencing higher education choices. Additionally, Villegas-Ch et al. [5] explored how machine learning models adapt educational content to match individual learning styles, leading to better learning outcomes. Advancements in smart education systems have also gained attention. Amin et al. [6] introduced an IoT-powered e-learning and MOOC recommender system designed to create adaptive learning experiences. Rajagopal et al. [7] examined different machine learning methods for online education, addressing key challenges in virtual learning environments. Meanwhile, Hukkeri and Goudar [8] developed a personalized recommendation system for e-learners, focusing on optimizing course selection based on student behavior. Beyond education, researchers have applied machine learning to personalized recommendations in other fields. Fang [9] explored a data-driven recommendation system that enhances tailored learning experiences. Qamhie et al. [10] designed a Personalized Career-Path Recommender System (PCRS) for engineering students, leveraging clustering and classification techniques to support career choices. Xiang and Zhang [11] proposed a fuzzy association-based approach to personalized recommendations in e-commerce, drawing parallels with education-focused recommendation models. Several studies have also examined course and content recommendation techniques using clustering and association rules. Asadi et al. [12] developed a hybrid course recommender system combining clustering and fuzzy association rules to personalize learning pathways. Zhou and Wang [13] proposed a data-mining-based English learning system that customizes exercises according to proficiency levels. Xu [14] explored computer-aided teaching recommendation systems, demonstrating how adaptive strategies

can enhance student performance. Additionally, Wang and Yang [15] studied the impact of deep learning in smart classrooms, highlighting improvements in personalized education for higher vocational training.

3 Proposed Methodology

3.1 Dataset

The dataset comprises of 2,000 student records that include personal, academic, and extracurricular information. The dataset contains the following attributes: first and last names, email, gender, part-time job status, number of absence days, extracurricular activities, weekly self-study hours, career aspirations, and academic scores in math, history, physics, chemistry, biology, English, and geography.

3.2 Model Workflow

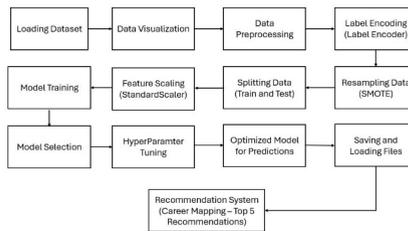


Fig. 1. Workflow diagram

Fig. 1 illustrates the workflow for developing a career recommendation system based on student data. The process begins with data visualization, where the dataset is analyzed to identify patterns, correlations, and potential issues such as missing values or class imbalances. Following this, the data undergoes preprocessing, where unnecessary columns are removed, and new features are added to improve predictive accuracy. Categorical variables are converted into numerical values using Label Encoding, and SMOTE (Synthetic Minority Over-sampling Technique) is applied to balance the dataset. The processed data is then split into training and testing sets, followed by feature scaling using StandardScaler to ensure consistency across numerical attributes. During model training, multiple algorithms are tested to identify the most effective one. Hyperparameter tuning is then performed to optimize performance. Once the best model is selected, it is saved for future use. Finally, the trained model is integrated into the recommendation system, which analyzes students' academic and extracurricular data to generate the top five career suggestions tailored to their profiles.

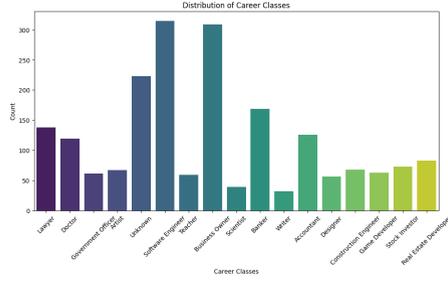


Fig. 2. Distribution of Career Classes

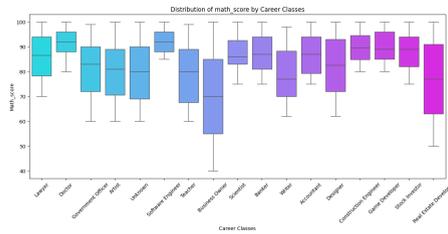


Fig. 3. Distribution of math-score by Career Classes

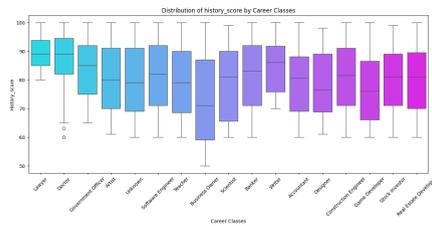


Fig. 4. Distribution of history-score by Career Classes

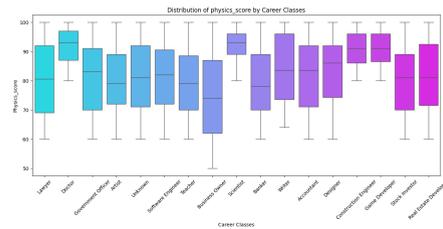


Fig. 5. Distribution of physics-score by Career Classes

4 Modules

Loading Dataset The dataset, obtained from Kaggle, includes 2,000 student records, each capturing academic performance and extracurricular involvement. It features subject-wise scores along with personal attributes and career aspirations. This data is then loaded into a Pandas DataFrame, where it undergoes preprocessing before being used for model training.

Data Visualization Data visualization plays a pivotal role in understanding the relationships between different variables. Figure 2 presents the distribution of career aspirations through a countplot, highlighting the frequency of each career class. Figures 3 to 8 showcase box plots for academic scores across various career classes, enabling the identification of outliers and trends within each class. These visualizations provide an initial understanding of the data structure and are essential for guiding feature engineering and model selection.

Data Preprocessing In this module, the unnecessary columns that do not contribute to predictions are removed, and new features are created based on domain knowledge. Additional attributes, such as total score and average score, are introduced to improve the model’s accuracy and predictive capabilities.

Label Encoding It converts categorical data into numerical values. This step is crucial because most machine learning algorithms require numerical input. By converting career categories into a numerical format, the model can effectively learn patterns and make accurate predictions, ensuring seamless integration into the training and evaluation process.

Feature Correlation Map To analyze how different features relate to each other, a correlation matrix is generated using Seaborn. This is visualized in Figure 9, where a heatmap highlights the strength of relationships between various feature pairs, helping identify patterns and dependencies within the data.

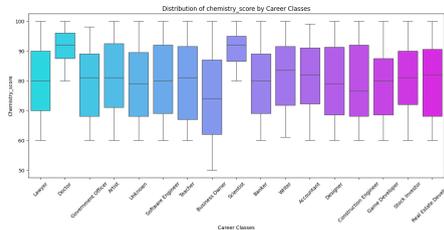


Fig. 6. Distribution of chemistry-score by Career Classes

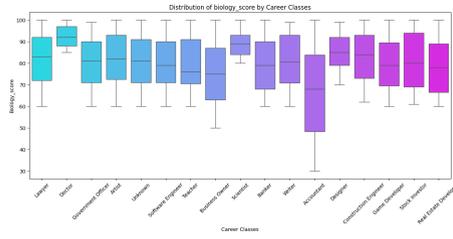


Fig. 7. Distribution of biology-score by Career Classes

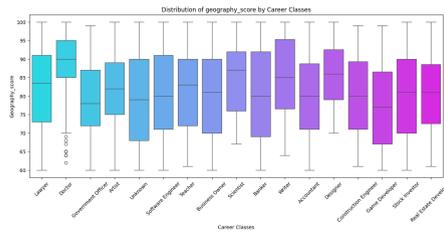


Fig. 8. Distribution of geography-score by Career Classes

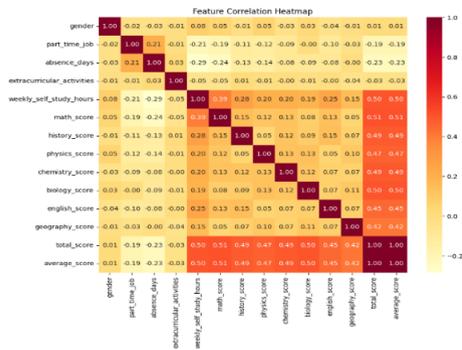


Fig. 9. Feature Correlation Heatmap

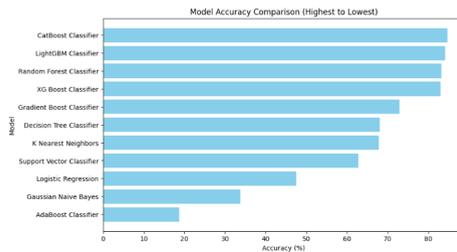


Fig. 10. Model Accuracy Comparison

```

Bayesian Optimized CatBoost Classifier Accuracy: 0.8571428571428571
Bayesian Optimized CatBoost Classifier Classification Report:
      precision    recall  f1-score   support

0         0.88      0.72      0.79         69
1         0.97      0.95      0.96         65
2         0.77      0.73      0.75         64
3         0.97      0.92      0.94         61
4         0.78      0.95      0.86         65
5         0.92      0.94      0.93         65
6         0.84      0.97      0.90         63
7         0.89      0.98      0.93         51
8         0.90      0.97      0.93         63
9         0.76      0.95      0.84         59
10        0.93      0.93      0.93         61
11        0.82      0.98      0.89         63
12        0.56      0.47      0.51         76
13        0.87      0.86      0.86         56
14        1.00      0.95      0.97         57
15        0.88      0.50      0.64         72
16        0.92      0.97      0.94         61

accuracy                0.86      1071
macro avg              0.86      0.87      0.86      1071
weighted avg          0.86      0.86      0.85      1071

```

Fig. 11. Classification Report of Bayesian Optimized CatBoost Classifier

Resampling Data To handle class imbalances that could impact model performance, SMOTE (Synthetic Minority Over-sampling Technique) is used. This technique creates synthetic samples for underrepresented career categories, helping the model learn more effectively from minority classes. By balancing the dataset, SMOTE reduces bias toward the majority class, leading to fairer and more accurate predictions.

Splitting Data into Train and Test Datasets The dataset is split into two subsets: training and testing. Eighty percent of the data is allocated for training, allowing the model to learn from a substantial portion of data. The remaining twenty percent is reserved for evaluating its performance. This separation ensures that the model is not just memorizing patterns but is capable of making accurate predictions on new inputs.

Feature Scaling To standardize the numerical features, StandardScaler is applied, which adjusts values by eliminating the mean and normalizing the variance to one. This transformation is particularly beneficial for algorithms that rely on distance-based calculations, as it ensures that no single feature dominates the learning process due to differences in scale. Moreover, in gradient-based models, such scaling expedites training by enhancing convergence speed, ultimately improving model efficiency.

Model Training A diverse set of machine learning models is implemented and assessed. Each algorithm is tested on the dataset, and their performance is measured using standard evaluation metrics. These models are selected based on

Welcome to the Career Mapping System!
 Please enter your gender (male/female): female
 Do you have a part-time job? (yes/no): no
 How many days have you been absent? 7
 Do you participate in extracurricular activities? (yes/no): yes
 How many hours do you study per week? 25
 Please provide your subject scores (out of 100):
 Math: 99
 History: 97
 Physics: 98
 Chemistry: 95
 Biology: 93
 English: 97
 Geography: 96

Top 5 Career Mapping Results

| Rank | Recommended Study | Probability |
|------|-------------------|-------------|
| 1 | Designer | 0.6484 |
| 2 | Teacher | 0.1272 |
| 3 | Banker | 0.1008 |
| 4 | Unknown | 0.0735 |
| 5 | Stock Investor | 0.0203 |

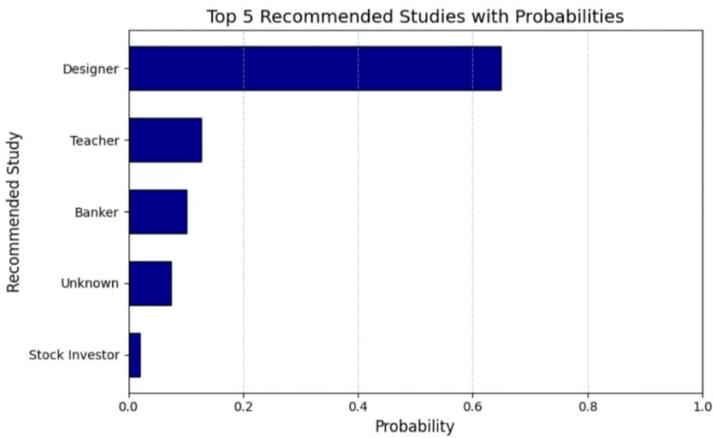


Fig. 12. Sample Output of Career Mapping System

their ability to handle mixed data types and their efficiency in solving classification problems, ensuring a robust approach to career prediction.

Model Selection On testing multiple models, CatBoost Classifier was the best with an accuracy of 84.69%. Known for its gradient boosting approach, CatBoost excels in handling categorical data and missing values, making it particularly well-suited for this dataset. Its strong performance makes it the preferred choice for career recommendations in this study. A comparison of model performances is presented in Figure 11, showcasing accuracy differences among the evaluated classifiers.

Hyperparameter Tuning To further improve the model, Bayesian Optimization is applied for hyperparameter fine-tuning. This approach efficiently explores different parameter combinations to identify the most effective settings for boosting performance. Key hyperparameters such as iterations, depth, learning rate, L2 leaf regularization, border count, subsample, and colsample by level are adjusted during the optimization process. As a result, the optimized model achieves an improved accuracy of 85.71% on the test dataset, demonstrating its enhanced ability to provide more precise career recommendations.

Optimized Model for Predictions The fine-tuned Bayesian CatBoost model is applied to generate career predictions tailored to each student's unique profile. By analyzing their academic performance and extracurricular involvement, the model offers well-informed career suggestions, aligning students with the most suitable career paths. This approach enhances the accuracy and reliability of career guidance, helping students make confident decisions about their future.

Recommendation System The Career Mapping System serves as the foundation of the recommendation engine, processing user inputs to generate personalized career suggestions. It presents the top five career paths, accompanied by probability scores and visualizations for better clarity. This approach ensures that students receive data-driven career insights, helping them make well-informed decisions aligned with their strengths and interests.

5 Results And Discussion

The Bayesian-optimized CatBoost Classifier was chosen as the final model for career recommendations, achieving an 85.71% accuracy. As shown in Figure 11, the classification report highlights strong precision and recall across most career categories. The Career Mapping System provides students with personalized career recommendations, presenting the top five career paths along with their corresponding probabilities. A sample system output, illustrating these recommendations, is depicted in Figure 12. This approach ensures data-driven, fair, and informed career guidance for students.

6 Conclusion

This machine learning-based career recommendation system automates career counseling by analyzing student profiles using classification models. After Bayesian optimization, CatBoost emerged as the best model, achieving 85.71% accuracy. The system generates top five career recommendations based on inputs like gender, study habits, subject scores, and extracurricular involvement, providing probabilities for each career path. This ensures consistent, scalable, and data-driven career guidance, minimizing manual intervention. By integrating this model, educational institutions can enhance personalized career counseling, helping students make informed career decisions. The model is also adaptable, allowing it to evolve with future educational advancements for a fair, accurate, and efficient recommendation process.

References

1. Guntupalli, U.G.S., Pandala, M.L., Veeranki, D.T., Kumbha, P.: Career Compass: A career path recommender using machine learning. In: IEEE International Conference on AI in Education (2024).
2. Siswipraptini, P.C., Warnars, H.L.H.S., Ramadhan, A., Budiharto, W.: Personalized career-path recommendation model for information technology students in Indonesia. *IEEE Transactions on Learning Technologies* 18(3), 112–124 (2024).
3. Bin, Q., Zuhairi, M.F., Morcos, J.: A comprehensive study on personalized learning recommendation in e-learning systems. *IEEE Access* 12, 34567–34578 (2024).
4. Kamal, N., Sarker, F., Rahman, A., Hossain, S., Mamun, K.A.: Recommender system in academic choices of higher education: A systematic review. *IEEE Transactions on Computers in Education* 19(2), 210–225 (2024).
5. Villegas-Ch, W., García-Ortiz, J., Sánchez-Viteri, S.: Personalization of learning: Machine learning models for adapting educational content to individual learning styles. In: IEEE Global Conference on Learning Technologies (2024).
6. Amin, S., Uddin, M.I., Mashwani, W.K., Alarood, A.A., Alzahrani, A., Alzahrani, A.O.: Developing a personalized e-learning and MOOC recommender system in IoT-enabled smart education. *IEEE Internet of Things Journal* 9(5), 8764–8775 (2023).
7. Rajagopal, M., Ali, B.M., Priya, S.S., Banu, W.A., Madhavi, M.G., Punamkumar: Machine learning methods for online education case. In: IEEE Conference on Smart Education Technologies (2023).
8. Hukkeri, G.S., Goudar, R.H.: Machine learning-based personalized recommendation system for e-learners. *Journal of Educational Computing Research* 58(4), 789–804 (2022).
9. Fang, Y.: Research on personalized recommendation system based on machine learning. In: IEEE International Conference on Data Science and Education (2022).
10. Qamhie, M., Sammaneh, H., Demaidi, M.N.: PCRS: Personalized career-path recommender system for engineering students. In: IEEE International Conference on Engineering Education (2020).
11. Xiang, D., Zhang, Z.: Cross-border e-commerce personalized recommendation based on fuzzy association specifications combined with complex preference model. *IEEE Access* 8, 34510–34521 (2020).

12. Asadi, S., Jafari, S., Shokrollahi, Z.: Developing a course recommender by combining clustering and fuzzy association rules. *Journal of Computer Education* 6(3), 289–305 (2019).
13. Zhou, L., Wang, C.: Research on recommendation of personalized exercises in English learning based on data mining. *IEEE Transactions on Educational Technology* 67(2), 145–160 (2021).
14. Xu, Y.: Computer-aided design of personalized recommendation in teaching system. *IEEE Computational Intelligence Magazine* 14(2), 34–45 (2019).
15. Wang, R.X., Yang, W.H.: The design and application of smart classroom teaching mode in higher vocational education based on deep learning. In: *IEEE Conference on Smart Learning Technologies* (2022).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

