Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

A Machine Learning-based Career Recommendation

Vaishnavi Nayak[1], Neha Vora[2]

[1]Student, [2]Assistant Professor, Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, India

Email: [1]nayakvaishnavi03@gmail.com, [2]nehavora2501@gmail.com

Abstract

Career selection is perhaps the most significant decision any student makes during their academic journey. This research thus presents a machine learning-based career recommendation system that will offer each student a career suggestion based on their academic performance and extracurricular involvement, including whether they hold a part-time job. Evaluations were conducted on several supervised machine learning models for predicting best career paths, such as Random Forest, Support Vector machine (SVM), and K-Nearest Neighbor (KNN) . Experiments revealed that Random Forest performed best and had an accuracy of 93%. The proposed system assists students in making informed career decisions based on data analysis.

Keywords: Career Selection, Random Forest, KNN, SVM.

1. Introduction

Career decisions are important during major transitions in a student's life, which, in turn, affects their future work and also their long-term job satisfaction, mental well-being, and career development. Such evolving factors call for the need to make important decisions when choosing the right career path among the numerous new businesses that spread fast. In general, career counseling has been provided through counseling, aptitude tests/evaluation, and psychometric assessments. While these methods are supportive, they have inherent limitations. These methods often overlook other important factors that affect a student's possible career, like academic strengths, personal interest, extracurricular involvement, and part-time job experience. Advancements in artificial intelligence, particularly in machine learning, are

DOI: https://doi.org/10.36548/jtcsst.2024.4.004
Received: 05.10.2024, received in revised form: 02.11.2024, accepted: 17.11.2024, published: 29.11.2024

paving the way for the development of automated career recommendations that offer highly personalized insights. These systems can perform millions of calculations to identify patterns in extensive datasets that are often beyond human recognition, enabling them to predict more suitable outcomes for individuals compared to traditional success programs aimed at helping someone achieve their goals. AI models can utilize the information about students' academic performance, extracurricular activities, personal traits, and declared career aspirations to suggest related careers that align with their skills and interests.

This study proposes a career recommendation system that leverages artificial intelligence to predict potential career paths for students based on a comprehensive analysis of personal and academic factors. To achieve this, several machine learning models are compared to identify the most effective computational approach for predicting career aspirations, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors. The study aims to build a system that gives students personalized career advice, helping students to explore options that match their skills and interests.

2. Related Work

Recently, the use of machine learning techniques in career recommendation systems has received tremendous attention. Yadalam in 2020, proposed a content-based filtering approach for career recommendations based on both academic performance and personal interests to match students with suitable opportunities [1]. Although effective, content-based filtering faces inherent limitations because content-based filtering is usually not able to include contextual data beyond that is considered in the content.

To overcome these constraints, Ezz and Elshenawy (2020) proposed an adaptive recommendation system that utilized a set of machine learning algorithms to predict the most suitable academic programs for students [2]. Their contribution emphasized that personalized systems if implemented into student advising, would further improve not just the preparation of the students for higher education but also help them develop professional skills among these students based on the paths suggested to them by their abilities and career objectives. The authors Wan and Ye(2022) then enhanced this by using deep learning models on college student career recommendations[3]. However, their work was later withdrawn for methodological errors. Despite the flaws in their methods, the study proved a possibility in the

use of deep learning techniques applied to large datasets can provide highly accurate career recommendations.

Al-Dossari (2020) identified the selection of the career path by using machine learning, mainly in the field of IT, where a student's ideal career choice would be determined based on a better grade, skills, and market demand[4]. More recent contributions from Santhosh, Shenoy, and Kumar (2023) present a machine learning-based system for student job role fit and career recommendation [5]. Their model grades students' academic performances, skillsets, and extracurricular activities to determine suitable jobs. It has been a key tool in giving broader general career recommendations that transcend mere academic performance.

The Babu and Mathew developed in 2023, was a two-tier system that used decision tree algorithms for career prediction and applied linear regression for the pass prediction, used especially on higher secondary school students who could be given individualized career advice according to academic trends and predicted academic outcomes [6] Wang made use of interpretable machine learning methods in mining campus big data to predict career choices for students. The system was supposed to bridge a gap between academic performance and the decisions at work, providing an interpretable way in which a career would be recommended[7].

It was observed that Zhou, Zhang, and Liu (2024) developed a system of AI-based career path recommendations based on the Myers-Briggs Type Indicator [8]. In that design, cross-cultural perspectives are accounted for so that career recommendations could correspond to personality types and cultural contexts, thus facilitating more personalized career advice. Siswipraptini et al. (2024) proposed a personalized model for information technology students in Indonesia for the recommendation of a career path [9]. The system was able to provide relevant career recommendations based on local industry requirements and profiles of student skills and aspirations.

Cui (2024) further extended this research by developing an integrated career interest assessment system for college students based on machine learning algorithms that enable the identification of various data from students and suggest applicable choices to them regarding career planning [10]. The main aim behind designing this system was to help students make that perfect choice of a career wherein they can best utilize their interests along with academic capabilities for assured long-term satisfaction in the chosen career.

In the review of job recommender systems, De Ruijt and Bhulai (2021) provided an overview of the existing systems of their strengths and weaknesses around the various approaches to machine learning that have dominated job recommendations [11]. Their efforts concentrated on what they said should be put in place over algorithmic transparency and interpretability for building trust with users.

Kamal, Sarker, and Mamun developed a comparative study of machine learning approaches for suggesting professors from different universities[15]. Their system employed measures for performance and academic profiles to match the expertise of professors with the requirements of any institution with some insights on how machine learning can enable streamlined matching of academia and professional. The CaPaR framework was developed by Patel, Kakuste, and Eirinaki. It is a career path recommendation system aimed at helping individuals in their career decisions based on an analysis of a customer's academic background and skills[12]. Their system adopted the use of techniques that combined both machine learning and data analysis techniques to provide specific, well-tailored career path recommendations based on individual profiles.

Majidi personalized course recommendations by developing a career goal-based system[13]. This system assisted the students in identifying the most relevant courses aligned with their long-term career aspirations, thus making the academic journey directly compatible with professional goals. Yet, as a third study, Jaber (2024) analyzed the way personality insights could be added to career guidance systems [14]. By incorporating personality insight models into machine learning analysis, this system can provide career suggestions aligned with students' academic and psychological profiles, offering a more integrated approach to career guidance.

3. Methodology

The methodology used here is a structured plan for developing and testing the career recommendation system based on machine learning. The dataset comprises 1,258 records of the students it includes. academic scores, gender, part-time job status, involvement in extracurricular activities, absence days, and weekly self-study hours are included in this dataset. There were 17 groups for career aspirations with 198 records. This approach involves data preprocessing, Exploratory Data Analysis (EDA), feature selection, model training, and model evaluation, each playing an essential role in ensuring the system's authenticity,

reliability, and effectiveness in providing personalized career guidance. The last important objective is building a robust model describing a student's career based on a combination of academic and personal attributes. By adhering to best practices in machine learning, the challenging steps ensure that models generalize well to new data, deliver meaningful recommendations, and address challenges such as class imbalances and data sparsity. Table 1 summarizes the distribution of career aspirations collected in the dataset, showcasing the range of career interests across different fields.

Table 1. Career Aspirations

Career Aspiration	**No. of samples collected**
Accountant	150
Software Engineer	200
Teacher	90
Lawyer	85
Scientist	100
Government Officer	110
Artist	50
Engineer	180
Doctor	120
Data Scientist	95
Architect	70
Writer	60
Entrepreneur	105
Real Estate Developer	50
Musician	40
Game Developer	75

The following questionnaires were used in collecting the students personal details.

Questionnaires

1. Have you participated in any extracurricular activities?
2. What are your weekly self-study hours?
3. What is your career aspiration(career goal)?
4. Provide your History score.
5. On average, how many days per month are you absent from school/college due to work or other reasons? (Please specify the number of days.)

This section provides a summary of the steps taken to develop the career recommendation system, such as Exploratory Data Analysis(EDA), feature selection, experimental setup, and model evaluation.

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to identify underlying patterns in the dataset. Exploratory Data Analysis (EDA) was used to find any kind of trends in both academic performance as well as study habits. Outliers were found in features like self-study hours and absence days and were removed from the analysis for better clarity, as some values, such as 0 to 22 absence days, were considered unrealistic. For most of the students, absence days ranged from 1 to 3. To smoothen this out and prevent any extreme outlier effect, it was capped at the 95th percentile.

The most important points were concluded as follows:

STEM Subjects: Academic achievements in Math, Physics, and Chemistry showed a good predictor of the relationship with technical careers.

Study Habits: The trend analysis of the hours of self-study done during the week and absence days showed some significant trends in the commitment of the students and its impacts on their performance.

Outliers Detection: The outliers were detected in study hours and absence days and hence were taken care of appropriately for the training of the model to make it unbiased.

These insights of deep things in turn decide which features are more related to describing career aspirations and contribute to the process of feature selection. Table 2 displays

the selected features from the Exploratory Data Analysis, outlining their importance in predicting career aspirations.

Table 2. Tabulated Features (Exploratory Data Analysis)

Feature	Type	Importance in Career Prediction
Mathematics Score	Numeric	Predictive of STEM-related careers
Physics Score	Numeric	Related to technical careers
Chemistry Score	Numeric	Vitals in science-related fields
Self-Study hours	Numeric	Display academic commitment
Extracurricular Activities	Categorical	Suggests leadership potential
Part-time Job	Categorical	Indicates real-world experience

Figure 1 illustrates the distribution of career aspirations based on the data analysis, highlighting the number of individuals aspiring to each career category.

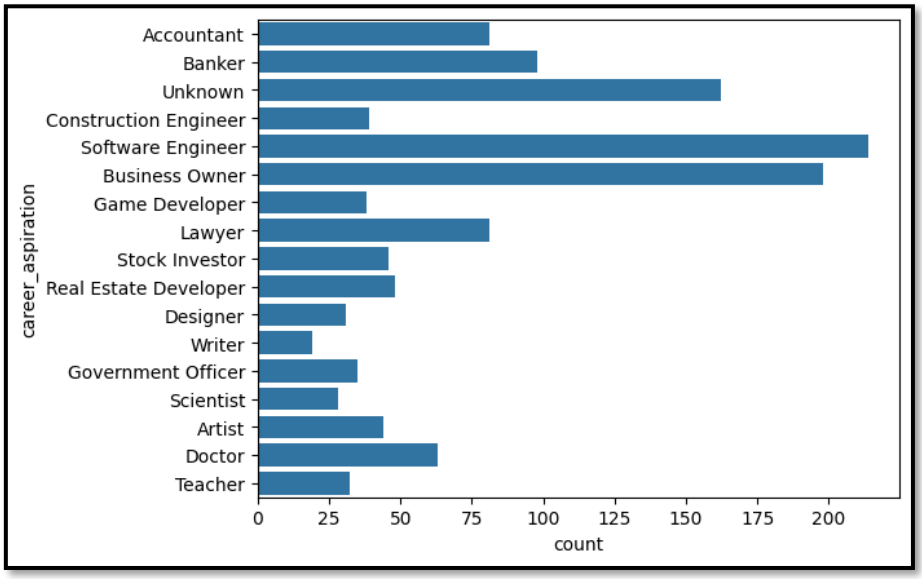

Figure 1. Distribution of Career Aspirations

The bar chart in Figure 1 illustrates the distribution of career aspirations, showing the number of people aiming for each type of career. The two most popular aspirations are Software Engineer and Data Scientist, with over 200 individuals aspiring to each. Other careers, such as Banker and Accountant, have fewer aspirants, while Teacher, Doctor, and Scientist are among

the least chosen options, as indicated by shorter bars. This EDA highlights a strong preference for careers in technology compared to other fields. Figure 2 presents a pair plot that visualizes the relationships between various academic and personal attributes, uncovering patterns linked to career aspirations.

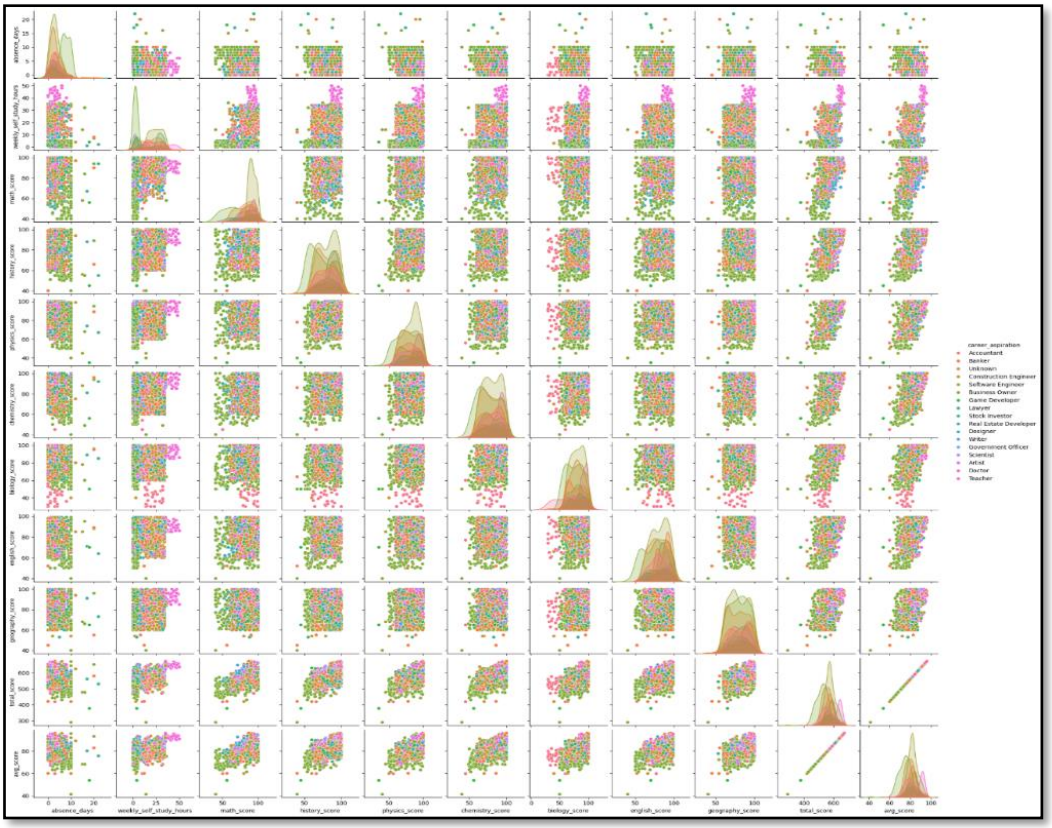

Figure 2. Pair plot

The pair plot in Figure 2 provides an in-depth exploratory analysis of various features in the dataset, including academic scores in subjects like Math, Physics, Chemistry, etc., study hours per week, and absences related to career aspirations. The color-coding for each career aspiration allows for visual comparison of correlations across different career paths, such as Accountant, Banker, and Software Engineer. Diagonal plots show the distribution of each feature, while scatter plots below the diagonal reveal potential relationships between pairs of features. Observed clusters suggest that certain career aspirations are strongly influenced by specific academic scores or study patterns. For example, aspiring to careers like Software Engineer or Data Scientist is associated with high scores in STEM subjects, indicating that strong performance in these areas is a good predictor of these careers. The plot also highlights variations in study habits, such as weekly study hours, across different career paths. Figure 3

presents a box plot that details the distribution of career aspirations, showing the central tendency and spread across various fields.

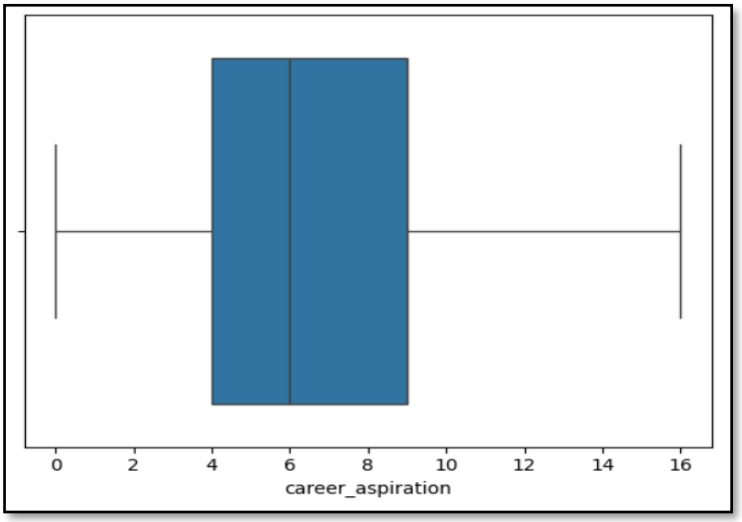

Figure 3. Box plot

The box plot in Figure.3 illustrates the distribution of career aspirations within different categories, with each category represented by a number. This type of box plot is useful for visually summarizing data based on the median, quartiles, and potential outliers. The middle 50% of the data is represented by the box, known as the interquartile range (IQR). A line within the box marks the median value. The "whiskers" extend to the minimum and maximum values, excluding outliers. In this plot, career aspirations range from 0 to approximately 16, with the median falling near the center of this range. No significant outliers are present, as all points fall within the whiskers, indicating a balanced distribution of career aspirations across categories

3.2 Feature Selection

Thus, Exploratory Data Analysis(EDA) leads to feature selection. Academic scores such, as grades in Mathematics, Physics, and Chemistry are the most important predictors of a career in STEM. Other than these, participation in extracurricular activities and part-time jobs have been selected to adjust for real-life experience and leadership potential.

Academic Scores: Mathematics, Physics, Chemistry, and Biology have been picked as major subjects since they are closely related to changes in their career path.

Self-Study Hours: Weekly self-study hours teach a student how loyal he/she can be to academics.

Absence Days: The absence days epitomized to what extent the students were involved. It was also made available to them the results concerning their career aspirations.

Extracurricular Activities & Part-Time Employment: We had included the non-academic characteristics to measure their influence on career choice, especially with factors like innovation and leadership.

3.3 Research Methodology

In the design of experiments, we have trained, tested, and assessed several machine learning models. The major emphasis was to measure how the models would perform in perceivable future career ambitions based on academics and other features.

The models applied in this study are selected based on being suitable for multi-class classification tasks, ease of implementation, and performance on academic and personal data. Three key models were chosen below:

Random Forest Classifier

Random forest is a robust method for multi-class classification and is strong against high-dimensional datasets. It is an ensemble learning method where average errors by multiple decision trees are used to reduce overfitting and improve generalization to new data. Additionally, it performs well with mixed feature types, including continuous and categorical. It also offers interpretability through feature importance.

Support Vector Machine (SVM)

SVMs can solve non-linear classification tasks using kernel functions. Since SVM is quite good at classification boundaries, mainly when it classifies based on a distinction between students' career aspirations using academic and personal attributes, SVM was used. SVM is very accurate for binary and multi-class classification issues provided the selected kernel is appropriate. Again, since the problems could be bound complexly, it would be suitable for dealing with numerous complicated decision boundaries.

K-Nearest Neighbors (KNN)

KNN is one of the most straightforward yet strong algorithms classified based on proximity. Its ease of understanding and its implementation were appropriate to be taken as a

baseline model to check for comparison with the more complex models. KNN works well with the small size of datasets and for non-parametric classification tasks. The features were scaled, and this was sensitive in KNN also, which was taken care of during the preprocessing stage.

4. Results and Discussion

Models were developed using precise tools and libraries. The research adopted Python as its programming language, which allows for the development and testing of machine-learning algorithms in a flexible setting. Among some of the most important libraries used in the study, was sci-kit-learn, which was used for building models like Random Forest, SVM and KNN and to support tasks such as model evaluation and cross-validation. All the data manipulation and preprocessing steps were performed using pandas and NumPy, which enabled fast cleaning, transformation, and feature selection of the data. Since visual results had to be generated, accuracy and loss curves and plot creation during EDA were done with the help of Matplotlib. The entire process, including training and testing, was conducted within a Jupyter Notebook.

The dataset was split into 80% for training and 20% for testing in a layered split to ensure that both subsets had a fair representation of all career categories. A 5-fold cross-validation was used to prevent overfitting and ensure the model well generalizes in different subsets of the data. The class of career aspirations in the dataset was imbalanced. This was addressed through the use of SMOTE, or Synthetic Minority Over-sampling Technique, to balance the training data. All missing data, most importantly the academic scores, were imputed using the mean.

Outlier detection and normalization were essential preprocessing steps before data could be used to train the model. Outliers for continuous variables like absent days and hours of self-study were detected and used at the 95th percentile. Predictions like scores and hours of self-study were normalized using Min-Max scaling to improve the distance-based models such as KNN. Attributes, like Extracurricular Participation and Part-Time Jobs, were one-hot encoded to represent them as two-class variables so that they can be used with machine learning models. Grid Search is used for Random Forest and SVM to get the best set of hyperparameters.

For the training models, some hyperparameters were fine-tuned to improve the classifiers' performances. For Random Forest, the no. of decision trees was increased to 200; the tree's max depth is set to 15 so it won't overfit the training data set. In SVM, the C regulation parameter was changed to 0.1, and it used the RBF kernel for dealing with non-linear features. The number of neighbors for KNN was set to 7 to smooth out decision boundaries. Some amount of early stopping is also done on the neural network to prevent overfitting, where training stops whenever validation accuracy stops improving. These have helped in balancing out accuracy and generalization among all the models.

Confusion Matrix: This was used to visually inspect the performance of each model in predicting different career classes.

The efficacy of the machine learning models was evaluated through a confusion matrix, a tool that gives a comprehensive detail of the entire set of predictions obtained from the model across different career aspiration categories. It is a breakdown of true positives, false positives, true negatives, and false negatives for each category. For example, the model got 140 as the number of correct predictions for the label Accountant, but it has been incorrectly labeled 5 as being a Software Engineer. Similarly, it got 190 instances of Software Engineer as correctly classified but has made 6 people's labels as that of an Accountant. The confusion matrix allows for evaluating the overall performance of the model in terms of precision, recall, and F1-score, among other metrics. This analysis helps determine how well the model generalizes across different career categories, its strengths in predicting certain aspirations, such as becoming a Software Engineer, and its weaknesses in distinguishing between careers with overlapping skill sets. Table 3 presents the confusion matrix used to evaluate the accuracy of the model in predicting different career aspirations. It outlines the correct and incorrect classifications across all career categories.

Table 3. Confusion Matrix for Career Aspirations

Actual/Predicted	Accountant	Engineer	Teacher	Scientist	Doctor	Lawyer	Total
Accountant	140	5	10	2	3	2	162
Engineer	4	485	7	5	6	3	210
Teacher	8	6	110	6	2	3	135

Scientist	5	7	6	95	4	1	118
Doctor	2	4	5	3	110	5	129
Lawyer	3	5	6	2	4	85	105
Total	162	212	144	113	129	99	859

The table shows the breakdown of correct and incorrect predictions for each career predictions.

Explanation

- True Positives(Diagonal elements): These represent the correct predictions for each career aspiration.
    - For instance, 140 accountants were correctly predicted as accountants, 185 engineers were predicted correctly, and so on.
- False Positives(Off-diagonal elements): These represent incorrect predictions.
    - For example, 5 engineers were misclassified as accountants, 7 scientists were misclassified as engineers, and so on.

Performance metrics Derived from the Confusion Matrix

Using this confusion matrix, the following performance scores for the random forest, SVM and KNN was calculated

- Accuracy: The overall correctness of the model.
- Precision: The number of correctly predicted instances out of all predicted instances for a class.
- Recall: The number of correctly predicted instances out of the actual instances for a class.
- F1-score: The harmonic mean of precision and recall.

Table 4 compares the overall results of the different machine learning models, highlighting the performance of each in terms of accuracy, precision, recall, and F1-score.

Table 4. Overall Results

Metrics/Model	Random Forest	SVM	KNN
Accuracy	93%	86%	73%
Precision	0.91	0.88	0.90
Recall	0.94	0.88	0.77
F1-score	0.93	0.86	0.83

Random Forest Classifier outperformed other models with the highest accuracy and balanced results for all career classes. SVM was another effective classifier, though it struggled with having too much data for some career aspirations and not enough for more specialized or less common ones KNN was the worst performer due to its sensitivity to scaling and its inability to handle multi-class problems effectively.

5. Limitation

Although the career recommendation system shows promising results, there are several limitations to consider. The dataset used consists of only 1,258 student records, which may limit the model's ability to generalize to larger and more diverse populations. Additionally, the system heavily relies on educational performance and extracurricular activities, while overlooking important factors such as psychological traits, personal motivations, and market trends, all of which can significantly influence career choices. Despite using SMOTE to address class imbalance, the system still faces challenges in predicting less common career paths with precision. Moreover, the simplification of career classification may overlook important distinctions between similar jobs. Finally, the system has not been tested in real-world scenarios, which could reveal practical limitations not apparent in controlled testing.

6. Conclusion

The development of a machine-learning-based career recommendation system marks a significant step toward providing personalized, data-driven career guidance for students. Using student data such as academic performance, extracurricular activities, part-time job status, and study habits, the system leverages the Random Forest classifier, achieving a notable accuracy

of 93%. This demonstrates the effectiveness of ensemble methods in handling complex datasets for career prediction. The model performed well in training and validation stages, showing increased accuracy, decreased loss, and strong generalization without overfitting. However, its current scope is limited to academic features, lacking consideration of psychological and motivational factors. Future enhancements could address these limitations by expanding the dataset, incorporating diverse educational backgrounds, integrating underrepresented career aspirations, and aligning recommendations with labor market trends. Real-world pilot testing and broader data inclusion could further improve its utility, offering students more comprehensive and future-oriented career guidance.

References

[1] Yadalam, Tanya V., Vaishnavi M. Gowda, Vanditha Shiva Kumar, Disha Girish, and M. Namratha. "Career recommendation systems using content based filtering." In 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India. IEEE, 2020. 660-665.

[2] Ezz, Mohamed, and Ayman Elshenawy. "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program." Education and Information Technologies 25, no. 4 (2020): 2733-2746.

[3] Wan, Qing, and Lin Ye. "[Retracted] Career Recommendation for College Students Based on Deep Learning and Machine Learning." Scientific Programming 2022, no. 1 (2022): 3437139.

[4] Al-Dossari, Hmood, Fawaz Abu Nughaymish, Ziyad Al-Qahtani, Mohammed Alkahlifah, and Asma Alqahtani. "A machine learning approach to career path choice for information technology graduates." Engineering, technology & applied science research 10, no. 6 (2020): 6589-6596.

[5] Santhosh, S., Ashwin Shenoy, and Sandeep Kumar. "Machine learning based ideal job role fit and career recommendation system." In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2023.64-67.

[6] Babu, Febin, P. G. Scholar, and Ms Meera Rose Mathew. "Career prediction using Decision tree algorithm and pass prediction using linear regression for higher secondary

school students." In Proceedings of the National Conference on Emerging Computer Applications (NCECA), vol. 5, no. 1, 404. 2023.

[7] Wang, Yuan, Liping Yang, Jun Wu, Zisheng Song, and Li Shi. "Mining campus big data: Prediction of career choice using interpretable machine learning method." Mathematics 10, no. 8 (2022): 1289.

[8] Zhou, Yizhou, Yong Zhang, and Naijie Liu. "Research on the Design of an AI Career Path Recommendation System Based on MBTI from a Cross-Cultural Perspective." Artificial Intelligence Technology Research 2, no. 1 (2024).

[9] Siswipraptini, Puji Catur, Harco Leslie Hendric Spits Warnars, Arief Ramadhan, and Widodo Budiharto. "Personalized Career-Path Recommendation Model for Information Technology Students in Indonesia." IEEE Access (2024).

[10] Cui, Can. "Career Interest Assessment: College Students Career Planning Based On Machine Leaning." Journal of Electrical Systems 20, no. 6s (2024): 1633-1644.

[11] De Ruijt, Corné, and Sandjai Bhulai. "Job recommender systems: A review." arXiv preprint arXiv:2111.13576 (2021).

[12] Patel, Bharat, Varun Kakuste, and Magdalini Eirinaki. "CaPaR: a career path recommendation framework." In 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), Redwood City, CA, USA pp. 23-30. IEEE, 2017.

[13] Majidi, Narges. "A personalized course recommendation system based on career goals." PhD diss., Memorial University of Newfoundland, 2018. https://research.library.mun.ca/13339/1/thesis.pdf

[14] Jaber, A. H. "Enhancing Career Guidance with Personality Insights: A Machine Learning Approach." PhD diss., ВНТУ, 2024. https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/41743/20938.pdf?sequence=3&isAllowed=y

[15] Kamal, Nabila, Farhana Sarker, and Khondaker A. Mamun. "A comparative study of machine learning approaches for recommending university faculty." In 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), IEEE, 2020. 1-6.

Author's biography

Neha Vora has completed Ph.D. in Computer Science and holds a Master's in Computer Applications (MCA). She is qualified in NET, SET, and GATE, and brings over 10 years of teaching experience, along with 1 year of industry experience. Her primary research areas include computer vision, image processing, machine learning, object detection, and artificial intelligence.