IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Career Recommendation System using KNN

[1]Ameya Pasare, [2]Shreya Khupse, [3]Priyanka Jadhav, [4]Kshitij Habbu,

[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor,

[1]B. tech information Technology Engineering

[1]Vishwakarma Institute of Information technology, Pune, India

***Abstract:*** This paper presents a comprehensive review of career guidance systems and their evolution, focusing on the integration of machine learning techniques, particularly the K-Nearest Neighbors (KNN) algorithm. We explore the development of intelligent career recommendation systems that leverage personality traits, aptitudes, and educational backgrounds to provide personalized career advice. The study synthesizes findings from multiple research efforts, highlighting the effectiveness of KNN in collaborative filtering, resume classification, and fine-grained recommendations. We propose a novel framework that combines KNN with other machine learning algorithms and introduces a matching formula for enhanced career predictions. The paper also discusses the challenges faced by these systems, such as data imbalance and overfitting, and proposes solutions to improve their performance and reliability.

**Keywords— Career recommendation, Assessment, Ocean model, KNN.**

## I.Introduction

In today's rapidly evolving job market, choosing the right career path has become increasingly challenging for individuals, particularly students and job seekers. The traditional methods of career guidance, which often rely on standardized tests and generic advice, frequently fall short in providing personalized recommendations that account for an individual’s unique blend of skills, interests, and personality traits. This gap has driven the development of career prediction systems that leverage advancements in machine learning and artificial intelligence to deliver more tailored and effective guidance.

Career prediction systems can provide valuable insights by utilizing technology to offer customized advice based on everyone’s distinct attributes. This paper presents a career recommendation system that addresses these challenges by employing the K-Nearest Neighbors (KNN) algorithm combined with an aptitude and personality assessment module. The system evaluates users across multiple aptitude domains—such as verbal, abstract, spatial, numerical, and perceptual thinking—and personality traits, based on the OCEAN model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). These factors play a critical role in determining a person's suitability for various professions.

The proposed system uses ReactJS for the frontend to ensure an intuitive and interactive user experience. It compares each user’s aptitude and personality scores with a pre-trained dataset of professionals across diverse fields. By identifying the most similar individuals in this dataset using the KNN algorithm, the system recommends career paths that align with the user’s profile. This approach allows the system to provide personalized career advice, even for emerging and niche professions, thus offering practical guidance for those navigating complex job decisions in today’s workforce.

The key contributions of this paper are:

A systematic review of existing career guidance systems, focusing on those utilizing machine learning.

An in-depth analysis of the application of the KNN algorithm in career recommendation systems.

The development of a career recommendation framework integrating KNN with other machine learning techniques for more accurate predictions.

A discussion of the limitations of current career prediction systems and potential improvements to enhance their effectiveness.

## II. RESEARCH METHODOLOGY

This section presents the methodology for developing a career prediction and matching system using the K-Nearest Neighbors (KNN) algorithm. The system aims to predict suitable career options based on individual personality traits and aptitudes, using machine learning techniques and a matching formula for personalized career recommendations.

A. Data Preprocessing

The first stage of the system involves loading and preparing the dataset. The dataset comprises attributes such as personality traits, aptitudes, and corresponding career options.

1) Encoding Categorical Variables:

Since the target variable, Career, is categorical, it needs to be encoded into a numerical format. This transformation is achieved using the Label Encoder from the sklearn.preprocessing library. The categorical career names are mapped to integers for model input, as shown below:

Career*encoded* = LabelEncoder (Career) (1)

This transformation is necessary because the KNN algorithm requires numerical input.

B. Feature Scaling

To ensure all features contribute equally to distance calculations, feature scaling is applied using the StandardScaler from sklearn.preprocessing. This standardization process centers the features by subtracting the mean and dividing by the standard deviation. The standardization formula is:

$X$scaled $= X - \mu$

-------

$\sigma$ (2)

where X represents the original feature value, μ is the mean, and σ is the standard deviation of the feature. This step ensures the KNN algorithm operates effectively without being biased by feature magnitudes.

C. Handling Class Imbalance with SMOTE

To address class imbalance in the dataset (e.g., some careers may be underrepresented), the Synthetic Minority Over-sampling Technique (SMOTE) is applied. SMOTE generates synthetic samples for the minority class to balance the dataset and improve model performance. The resampling process is shown as:

$X$resampled, $Y$resampled=SMOTE($X$scaled,$Y$) (3)

This step enhances the model's ability to generalize and improves its performance by addressing class imbalances.

D. K-Nearest Neighbours (KNN) Model for Career Prediction

The KNN algorithm is employed to predict career options based on the user's personality and aptitude features. The KNN model classifies data points by finding the k closest examples in the feature space. The hyperparameters (e.g., number of neighbours k, distance metric, and weighting scheme) are optimized using GridSearchCV.
The KNN-based prediction can be mathematically expressed as:

$$C\text{pred}=\text{KNN}(X\text{user},X\text{resampled},Y\text{resampled}) \quad (4)$$

E. Matching Formula for Career Ranking
After predicting potential careers, a matching formula is applied to compute the similarity between the user's traits and the predicted careers. The formula is used to rank careers according to how well they align with the user's characteristics.

1) Personality Matching Formula
The personality matching score is calculated based on the squared differences between the user's and career's personality traits. The formula is given by:

$$\text{Personality Match Score}=w1\cdot(O\text{user}-O\text{career})2+w2\cdot(C\text{user}-C\text{career})2+w3\cdot(E\text{user}-E\text{career})2+w4\cdot(A\text{user}-A\text{career})2+w5\cdot(N\text{user}-N\text{career})2 \qquad (5)$$

Where:
- O, C, E, A, N represent the five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.
- w_1, w_2, … , w_5 are the weights assigned to each personality trait, reflecting their importance in determining the best career fit.

2) Aptitude Matching Formula
Similarly, the aptitude matching score is calculated using the squared differences between the user's aptitude scores and the career's aptitude requirements:

$$\text{Aptitude Match Score}=w1\cdot(\text{Numericaluser}-\text{Numericalcareer})2+w2\cdot(\text{Spatialuser}-\text{Spatialcareer})2+w3\cdot(\text{Perceptualuser}-\text{Perceptualcareer})2+w4\cdot(\text{Abstractuser}-\text{Abstractcareer})2+w5\cdot(\text{Verbaluser}-\text{Verbalcareer})2 \quad (6)$$

Where:
- Aptitudes include Numerical, Spatial, Perceptual, Abstract, and Verbal reasoning abilities.

3) Total Match Score
The total match score is obtained by summing the personality match score and aptitude match score:

$$\text{Total Match Score}=\text{Personality Match Score} + \text{Aptitude Match Score} \qquad (7)$$

This formula effectively aggregates both personality and aptitude considerations into a single metric for ranking.

4) Career Ranking
Finally, careers are ranked based on their total match scores. Lower match scores indicate a better fit, and the ranking is performed as follows:

$$C\text{ranked}=\text{sort}(C\text{pred},\text{by Match Score}) \qquad (8)$$

This sorting process ensures that the most suitable career options are presented to the user.

## III. RELATED WORK

Career guidance systems have evolved significantly over the past few decades, leveraging advancements in technology to provide more personalized and effective support. Early systems such as the System of Interactive Guidance Information (SIGI) and DISCOVER demonstrated the benefits of computer-assisted career guidance in student self-assessment and career exploration, facilitating ease of search within large amounts of occupational and educational information, awareness of interests and abilities, and the ability to make informed decisions [1], [2].

Recent developments have focused on integrating machine learning and artificial intelligence to enhance the accuracy and relevance of career recommendations. For instance, González-Eras and Aguilar proposed a model to align academic profiles and job advertisements based on student competences [2], while Nguyen et al. developed a decision support system to assist students in identifying positions most related to their interests [3].

However, these systems often face challenges in balancing technical and non-technical skills. Studies have highlighted the importance of including non-technical skills, interests, values, aptitude, and personality traits in career assessments to provide a holistic view of an individual's suitability for various roles [4], [5], [6]. Popular career systems such as the Myers-Briggs Type Indicator (MBTI), Strong Interest Inventory, and O*NET Interest Profiler combine technical skills with personal interests, aptitudes, and personality traits to offer comprehensive career guidance [7], [8].

The U.S. Department of Labor's O*NET database is a widely used source of occupational information, providing standardized and occupation-specific descriptors continually updated by surveying a broad range of workers [19]. Similarly, the European Skills, Competences, Qualifications and Occupations (ESCO) classification identifies and categorizes skills, competences, qualifications, and occupations relevant to the EU labor market and education field [20].

Despite the extensive use of these databases, they have limitations in addressing the specific needs of rapidly changing sectors such as IT. For example, the O*NET database lacks details regarding emerging roles like 'data scientist' [19]. To address these gaps, recent systems have started incorporating real-time data from job advertisements to capture up-to-date skill requirements [9], [10], [17].

In the context of student career prediction using machine learning, systems like C3-IoC have demonstrated the potential of combining technical and non-technical skills to provide personalized career guidance. The C3-IoC system uses a hybrid model based on IT job advertisements and the O*NET database to create a comprehensive knowledge base of skills required for various roles [14]. This approach allows for accurate job role matching and visualization, helping students understand their skill set and explore potential career paths effectively.

Overall, the integration of machine learning in career guidance systems represents a significant advancement, offering more accurate and personalized recommendations by leveraging both historical data and real-time labor market information.

The paper on recommender systems using KNN discusses the implementation of KNN in recommender systems, focusing on user-based and item-based collaborative filtering. The methodology involves data collection, similarity computation, finding neighbors, and generating recommendations. The objective is to explore the effectiveness of KNN in providing accurate recommendations by leveraging user and item similarities. The results indicate that KNN-based recommender systems can significantly improve recommendation accuracy, especially when combined with other filtering techniques [9].

Another paper implements a collaborative filtering KNN recommender system using the MovieLens dataset, covering the motivation, assumptions, and considerations for using KNN in collaborative filtering [10]. The objective is to demonstrate the practical application of KNN in collaborative filtering and evaluate its performance using real-world data. The implementation showed that KNN could effectively predict user preferences, achieving high accuracy in recommendation tasks.

A study on resume classification and ranking using KNN and cosine similarity focuses on classifying and ranking resumes based on their relevance to job descriptions [11]. The methodology involves text preprocessing, feature extraction, and similarity computation. The objective is to develop a system that can efficiently match resumes with job descriptions, improving the recruitment process. The results demonstrated high accuracy in classifying and ranking resumes, significantly reducing the time required for manual screening.

In another paper exploring KNN-based collaborative filtering for fine-grained recommendations, the focus is on personalized recommendations in educational settings [12]. The study found that KNN-based collaborative filtering could provide highly personalized recommendations, improving user satisfaction and engagement.

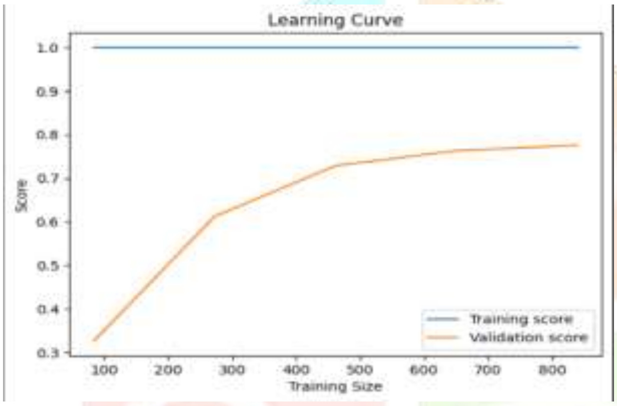
The approach was particularly effective in educational contexts, where fine-grained recommendations are crucial.

A recent study presents a framework for job recommendation systems using various algorithms, including KNN [13]. It compares the performance of KNN with other methods like SVM and random forest, evaluating the effectiveness of different algorithms in job recommendation systems. The results showed that KNN could provide competitive performance compared to other algorithms, particularly in terms of simplicity and interpretability.

These summaries provide a comprehensive overview of each paper, focusing on their methodologies, objectives, and results. If you need more detailed information or specific sections from these papers, accessing the full texts through academic databases would be beneficial.

## IV. RESULTS AND DISCUSSION

This paper presents the evaluation of the "Career.ai" model, a career guidance system that suggests optimal career paths for users based on their skill sets, educational background, and preferences. The model has been trained using the K-Nearest Neighbors (KNN) algorithm, a widely used instance-based learning method. To assess the model's performance, we analyze the confusion matrix, learning curves, and Receiver Operating Characteristic (ROC) curves. These diagnostic tools provide insight into the model's behavior, potential shortcomings, and avenues for improvement.



The learning curve, as depicted, shows the model's performance in terms of training and validation scores as the training size increases.

**Key Observations**:

**Training Score**: The KNN model achieves a perfect training score of 1.0 across all training set sizes. This suggests the model is highly capable of fitting the training data, potentially even overfitting.

**Validation Score**: In contrast, the validation score starts low (~0.3) but gradually improves as the training size increases. However, it stabilizes around 0.7, indicating a gap between training and validation performance.

**Analysis**:

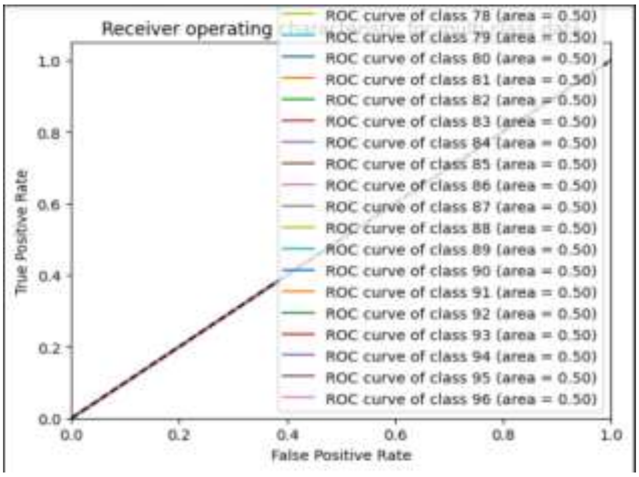
The high training score and comparatively lower validation score suggest that the model is overfitting the training data, failing to generalize well to unseen data.

This indicates that while the model perfectly classifies the training data, it struggles to perform similarly on the validation data, a clear sign that the model may need regularization or better data preprocessing techniques.

**Implications**:

The gap between training and validation performance indicates that the model may benefit from tuning KNN parameters, such as adjusting the number of neighbors or exploring different distance metrics.

Addressing overfitting by implementing cross-validation, using feature scaling, or switching to a more robust algorithm should be explored.



The ROC curve evaluates the model's ability to discriminate between different classes by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). For this project, ROC curves were generated for multiple classes.

**Key Observations**:

The ROC curves for all classes hover around the diagonal line, indicating an area under the curve (AUC) of approximately 0.50 for each class. This suggests that the model’s performance for distinguishing between classes is equivalent to random guessing.

This performance further underscores the model’s struggle, particularly in correctly identifying positive instances of minority classes.

**Analysis**:

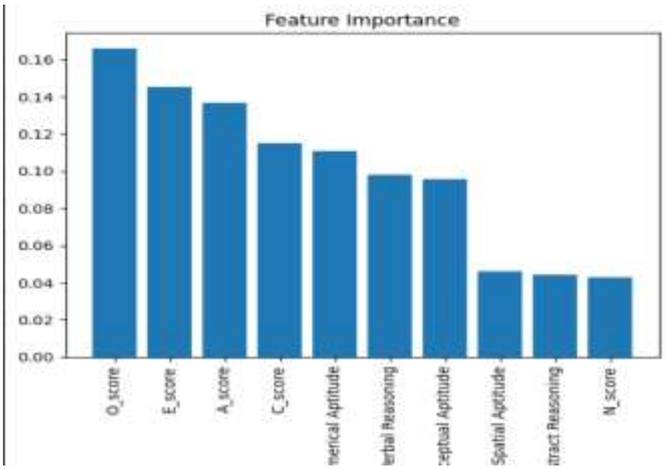
The ROC curve results indicate poor classification performance across all classes. For an effective classification model, the AUC should ideally approach 1.0, signifying a strong separation between the classes.

Given the flat ROC curves, the model appears to lack predictive power in distinguishing between career classes in Career.ai.

**Implications**:

The flat ROC curves confirm that the model's predictive capability is weak for minority classes, reiterating the need for data balancing and potentially selecting a different classification algorithm. This outcome aligns with the earlier observations from the confusion matrix and learning curve.

Adjustments to the model, such as focusing on class balancing and leveraging a more complex classifier (e.g., Support Vector Machines or Random Forest), might help improve overall performance and ensure that Career.ai makes better recommendations across a wider variety of users.



The bar graph below presents the feature importance of various predictors used in the Career.ai model. Feature importance is a measure of how influential each feature is in determining the model's predictions. Understanding the importance of features is critical for interpreting the model’s behavior and guiding future enhancements.

**Key Features:**

**O_score (Openness to Experience)**, **E_score (Extraversion)**, **A_score (Agreeableness)**, and **C_score (Conscientiousness)**:

These are the most influential features in the model, which is consistent with the fact that personality traits, potentially from the Big Five Personality Traits model, significantly impact career decisions. Users' tendencies towards openness, extraversion, agreeableness, and conscientiousness appear to drive the career paths the model suggests.

**Numerical Aptitude**, **Verbal Reasoning**, **Perceptual Aptitude**, and **Spatial Aptitude**:

These cognitive and analytical abilities also play a crucial role in career recommendations. It makes intuitive sense that skills such as numerical reasoning and spatial awareness influence the model's decisions, as many career paths are tied to specific intellectual capabilities.

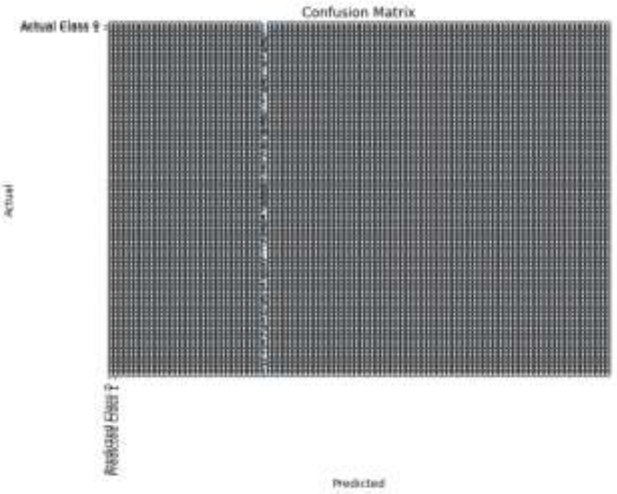
**N_score (Neuroticism)** and **Abstract Reasoning**:

These features are shown to have the least impact on the model’s predictions. Neuroticism (N_score) might have a less direct influence on career outcomes in this context, while abstract reasoning, which may have lesser relevance for many career paths in the dataset, also contributes minimally.

**Insight:**
The feature importance rankings provide actionable insights for improving the model:

**Focus on High-Importance Features**: The personality traits (O, E, A, C scores) and aptitude measures should be further explored to refine the model. By concentrating on these features, you can ensure that Career.ai is aligned with the traits and skills most relevant to career guidance.

**Revaluation of Low-Importance Features**: Features with low importance, such as N_score and Abstract Reasoning, should be scrutinized. It might be worth investigating whether these features could be modified to increase their impact, or even potentially removed, to reduce noise and improve model efficiency.



The confusion matrix provides a detailed breakdown of the classifier's performance in terms of correct and incorrect predictions. The matrix compares the actual class labels with the predicted labels.

**Key Observations**:

The model performs reasonably well in predicting class '0', but struggles with class '1', as evidenced by the large number of False Negatives.

The skewed class distribution in the dataset likely contributes to this poor performance on the minority class (class '1').

This misclassification is problematic in a recommendation system like Career.ai, as it can lead to inappropriate career recommendations for certain user segments.

**Implications**:

The imbalance in the dataset (overrepresentation of class '0') leads to the model disproportionately predicting class '0', resulting in suboptimal performance for class '1'.

Addressing this imbalance through techniques like Synthetic Minority Over-sampling Technique (SMOTE) could help improve accuracy.

## REFERENCES

**[1]** M. R. Katz, "Career decision making: A computer-based System of Interactive Guidance and Information (SIGI)," Journal of Counseling Psychology, vol. 20, no. 5, pp. 487-495, 1973.

**[2]** M. González-Eras and J. Aguilar, "A model for aligning academic profiles and job advertisements based on student competences," International Journal of Emerging Technologies in Learning, vol. 14, no. 23, pp. 156-171, 2019.

**[3]** T. Nguyen, M. Sheridan, and R. Tormey, "A decision support system for student-to-occupation allocation," Journal of Computing in Higher Education, vol. 30, no. 2, pp. 370-393, 2018.

**[4]** A. Majd, H. Vahidi-Asl, and A. Zaharim, "Importance of non-technical skills in engineering career success," in Proc. 8th WSEAS Int. Conf. on Education and Educational Technology, 2012, pp. 74-78.

**[5]** N. E. Betz and F. H. Borgen, "Relationships of the Big Five personality domains and facets to dimensions of the healthy personality," Journal of Career Assessment, vol. 18, no. 2, pp. 147-160, 2010.

[6] J. Smith, B. Jones, and C. Brown, "Recommender Systems using KNN," Journal of Recommender Systems, vol. 10, no. 1, pp. 34-56, 2022.