MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 2s (2026) | Pages: 178-190

A Hybrid Machine Learning Approach for Career Prediction Using Academic Features and Psychometric Personality Models

Nidhee Mandle[1], Nupur Thakre[2], Priyanka Shendage[3], Ramkrishna Vadali[4], Shradha Tawade[5]

[1,2,3,4,5]Information Technology, Pimpri Chinchwad College of Engineering, Pune,Maharashtra, India-411044

[1]nidhee.mandle22@pccoepune.org, [2]nupur.thakre22@pccoepune.org, [3]priyanka.shendage22@pccoepune.org, [4]ramkrishna.vadali@pccoepune.org,[5] shradha.tawade@pccoepune.org

Peer Review Information

***Type**: Article*
***Received**: 27 March 2026*
***Revised**: 12 April 2026*
***Accepted**: 26 May 2026*
***Published**: 16 June 2026*

Abstract

Selecting an adequate career requires taking into account various elements including academics, individual interest areas, and personality characteristics. Conventional career recommender models usually consider only one of these aspects, which is either academics or interest profiling, yielding relatively poor accuracy rates. This paper explores a multi-modal machine learning paradigm that utilizes academic characteristics, vocational interests, and personality traits in recommending careers in order to improve recommendation efficiency and robustness. Three major types of datasets will be applied for training: a Career Recommender Dataset with academic features, a massive RIASEC dataset that provides information about interests, and the Big Five personality dataset. Academic features will be normalized with TF-IDF and NLP techniques, and a logistic regression model will be first generated based on this data. It will then be improved by the addition of psychometric features, and a stacked ensemble will be constructed where a Random Forest meta-classifier will make final recommendations. Experimental analysis will reveal that the integration of diverse sources yields substantial prediction performance advantages in contrast to the use of academic-only data, and it may even provide valuable insight regarding personality-career associations.

Keywords: Career Recommendation System; Stacked Ensemble Learning; RIASEC Vocational Interest Model; Big Five Personality Traits; Natural Language Processing (TF-IDF); Machine Learning for Career Guidance.

How to Cite This Article

Mandle, N., Thakre, N., Shendage, P., Vadali, R., & Tawade, S. (2026). A hybrid machine learning approach for career prediction using academic features and psychometric personality models. *Multidisciplinary Journal of Research in Engineering and Technology*, 13(2s), 178–190.

Introduction

The career choice-making process can be complicated due to the many different aspects that affect it -- e.g., academic history, interests, skill set, and personality traits. With the expanding range of available educational opportunities and career options, students may find it challenging to find a career path that uses their strengths and preferences. Most traditional methods of providing career guidance are based on personal interviews, physical evaluations, or separate personality tests, which do not adequately represent the many dimensions of someone's profile. New, advanced data processing and machine learning technologies have created the capacity for the establishment of intelligent career recommendation systems that will use large amounts of data to provide tailored career recommendations based on a number of different factors.

Nowadays with all of the challenges to people looking for jobs around the world, there are many more ways to recommend jobs based on large amounts of data using more computational techniques. A new generation of career recommendation systems will be based on a much broader, complete picture of how different areas of school and work are working together.The research part of this project leverages machine learning algorithms to build an automated recommendation engine combining several forms of data together. We will use Python libraries such as Pandas and NumPy for our base level of data modifications; Scikit-learn will provide machine learning capabilities; and finally, we will use various NLP (Natural Language Processing) techniques (e.g., TF-IDF – Term Frequency-Inverse Document Frequency) to create numeric vectors that represent each term contained within a text document so that they can be analyzed through clustering (i.e., K-means) and classification (Logistic Regression, Random Forest) algorithms.

This paper outlines a framework that employs three separate analytic perspectives in order to establish a higher accuracy of career recommendations. The first perspective entails examining the career recommendation data set through an analysis of the academic and skills-based aspects of individuals' educational backgrounds. The second perspective uses the RIASEC model to capture the six psychological dimensions of career testing – Realistic, Investigative, Artistic, Social, Enterprising and Conventional. The third perspective incorporates the Big Five model as a representation of personality traits: extroversion, agreeableness, openness, conscientiousness and neuroticism. The use of stacked ensemble learning to combine and utilize data from the three different perspectives to provide a single recommendation allows the rankings to be based upon objective measurements and the measurement of psychological characteristics.

This research has shown that by combining academic indicator data with psychological models of personality and integrating these sources of information into a single collection of data, we can get a better picture of an individual's fit for a specific career. When we use multiple data sources, the system can find relationships or patterns between the data sources that might otherwise remain unrecognized if only one data source was collected or used. The results obtained from experimental evaluation indicate that the ensemble-based meta model is able to produce valid predictions of career fit across many career fields while also enhancing interpretability of how an individual's personality traits and vocational interest contribute toward the overall fit of individuals with their careers.

While there are plenty of career guidance resources available today, many of the current systems are still reliant on only a few pieces of information, like a student’s grades in school or a simple standardized test of aptitude, which leaves users with an incomplete and less personalized recommendation based solely on their academic history. The problem that this research addresses is that there are very few examples of integrated career recommendation models that [now] simultaneously take into consideration a person’s academic credentials; vocational interests and workplace personality type — using valid data analytics. This research will fill this gap by developing a multi-modal machine learning approach that can merge these three distinct data sources into one predictive model for recommendation; thereby allowing students and entry-level professionals to obtain more reliable and relevant career recommendations.

Literature Survey

Machine learning (ML) and artificial intelligence (AI) are drastically changing the ways in which individuals employ automated career/professional counseling. Traditional career/professional counseling systems use human-based analytical and evaluative tools, which have been ineffective and are lacking the ability to adapt to new labor market conditions.

In addition to developing ML classifiers and using ensemble learning models to generate optimal career options based on an individual’s education, skills, and interests, many advancements have occurred in the field. An example is a study done by Faruque et al. [1] that used NLP and machine learning as a means of predicting a computer science career for students; this was done to assess how well ML-based methods could provide viable career options to students. Similarly, Guru et al. [2] demonstrated that profiles of individual personality types could be integrated with profiles of skills through ML-generated recommendations for personalized career development (i.e., suggestion of how best to utilize one’s skills and personality together to determine a course of action in choosing a career).

Utilizing resume datasets like the Kaggle Resume Dataset is beneficial for recommending career options, as multiple studies in prior research have utilized resumes for the purpose of providing career and employment recommendations. As an example of this, Reddy et al. [4] developed a Resume Analyzer tool which uses Natural Language Processing (NLP) to analyze the skills and qualifications of

applicants to provide job recommendations based on their resumes in Reddy's Resume Database. Next, Kulkarni et al. [8] added to Reddy's system by combining different Deep Learning Algorithms to perform an analysis of the semantic properties of other resumes in the analysis data set. This current study confirms that investigating and using resume data sets like the Kaggle Resume Data Set are effective tools in supporting readers' research in choosing a career.

There is ongoing growth in the interest of creating resources to help students make educated career decisions based on student-related activities along with resumes. To develop these types of resources, Sylvia et al. created a smart advisor to provide recommendations on prospective students' careers by using machine learning to connect students' academic performance and their actions as students with their capabilities and limitations as a potential professional, an idea presented by Sylvia et al. Taller et al. [4] used this framework to provide students with career recommendations using learning analytics and career counselling to consider behaviours and any extra-curricular characteristics from the combined results of the suggested survey study to assist students who want to explore careers earlier.

Therefore, behavioral and psychometric analysis have become critical aspects of contemporary career guidance systems' recommendations. Shahzada et al. (2020) developed CAREERLLAMA – a methodology that integrates AI-based advice, skills deficiency analysis and psychometric data, providing students with customized career advice. Furthermore, Singh Sisodiya et al.'s (2020) dual approach to combining technical and personality characteristics of individuals into a single instrument for personalizing career advice has demonstrated an increase in accuracy over past personalized career advice methods. This data supports psychometric testing as part of an optimum recommendation system for students aged 13 to 14.

There have been positive outcomes in utilizing hybrid recommendation methods formed through combining collaborative filtering with content-based filtering. Ong and Lim proposed a skill recommendation architecture based on the customer's profile being linked to the skills that relate to the industry [11]. Wang et al. developed an AI-driven career discovery model in which recommendations are adapted to the user profile based on user actions in an active manner [10]. A key feature of this type of system is the design and implementation of self-learning systems.

Automation and learning processes have been heavily utilized by career suggestion systems. AutoML-based methods allow models to be automatically updated and optimized with flexibility in response to data and user behavior, according to a number of recent literature reviews [14], [17]. Mujtaba and Mahapatra addressed a number of ethical concerns regarding AI-based recruiting and career advice systems, including fairness and prejudice in AI applications [18].

The literature clearly demonstrates a shift toward the provision of AI-driven, individualized, flexible career advice systems, which use various types of technology such as resume processing, behavioral assessment, psychological assessment, and automated learning pathways. The number of systems currently in existence that will successfully operate across multiple age groups and update themselves automatically is very limited. The Self-Enhancing AI-Powered Career Recommendation System is proposed to utilize a large variety of data sources, advanced machine learning models and an AI agent to improve the overall performance of the system, thereby addressing these gaps.

Data Set

This research describes a new machine learning-based career recommendation system developed using three distinct types of datasets representing the various dimensions of career decision-making, including educational background, vocational interest, and personality traits. The first dataset to be used for this study is called the Career Recommendation Dataset. The Career Recommendation Dataset contains records relating to students' educational background, academic performance and achievement (e.g., GPA, grade level), interest in careers (e.g., hobbies), strengths and weaknesses in certain areas of work, certifications completed, career goals, job titles, and finally, career outcomes. The Career Recommendation Dataset is the main dataset used as the basis for classifying different occupations into Related Fields or 8 major categories (Core Engineering, Data Science, IT Software, Finance & Business, Marketing & Sales, Healthcare & Legal, Education, and Other). All text attributes are processed using TF-IDF vectorization (natural language processing) to convert unstructured text data into machine-readable feature vectors, which can then be used to build a machine learning model.

Two additional large-scale psychological datasets have been combined with academic data to analyze personality and vocational interest behaviors. The RIASEC Vocational Interest Dataset contains survey responses from a personality assessment using the RIASEC model (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional). Over 145,000 respondents provided survey responses allowing for the extraction of vocational interest scores for identifying individuals' preferred working environments.The second additional large-scale psychological dataset is the Big Five Personality Dataset which contains responses from over 500,000 subjects collected through an online personality assessment using the OCEAN model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). The Big Five traits are computed from questionnaire responses, therefore providing stable personality characteristics that will influence an individual's fit with a career. The combination of three datasets provides academic indicators, vocational interests, and personality traits

for multidimensional analysis for more appropriate and personalized career recommendations.

Proposed Methodology

A new system is being created to help people make career decisions through a computerized recommendation process using multi-modal machine learning techniques. This system will consider multiple factors in combining academic, vocational and personality traits for better career recommendations than traditional systems which only consider one of these categories (i.e., academic performance or abilities). The base of data used in this system will include three independent sources of information: 1) Academic qualifications and education history, 2) RIASEC vocational interests and propensity toward each of the six RIASEC styles (Realistic, Investigative, Artistic, Social, Enterprising and Conventional), and 3) Big Five personality traits (Openness to Experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism). Therefore, the goal is to develop a robust and realistic predictive model that has the ability to understand the different ways these three categories interact to inform an individual's career suitability. The proposed methodology will employ several emerging technologies such as natural language processing (NLP), additional techniques such as feature engineering and clustering will also be employed as part of developing the full prediction model. Therefore, through the application of NLP, additional applications will be able to 'capture' and integrate heterogeneous data sets into a single prediction model.

The proposed system's architecture consists of three main processing layers followed by a meta-learning layer. The first layer includes the processing of academic records to provide predictions about possible career fields based on both textual & numeric information. In layer two, we will extract vocational interest using the RIASEC personality framework to determine how well the student will fit into a particular career (if they were working in that field). In layer three, personality characteristics are determined using the Big Five (OCEAN) Personality Model. The outputs of all three layers are then combined using a stacked ensemble architecture with one meta-classifier that creates a combination of predictions from the individual components as its final recommendation of what career the student should pursue.

System Architecture

After the extraction of the feature sets for each dataset, the vectors are combined into a single feature representation. The final architecture that is proposed has four basic layers. The first layer is data acquisition; the second layer is data preprocessing and feature extraction. The third layer is base model processing, and the fourth layer is meta-level prediction. The first layer consists of three different sets of independent data obtained from different points of view in the selection of jobs. The sets are: academic characteristics (from the Career Recommendation Database), vocational characteristics (from the RIASEC Database), and personal characteristics (from the Five-Factor Model Database).

Before being used in analyses, the datasets go through preprocessing, a stage where they are cleaned, filtered, and transformed to remove inconsistencies and prepare them for analysis. Features of the academic text are their respective skills, specialization, certification, and interest, which all combine and are processed at this stage using the TF-IDF vectorization technique as a method of converting the unstructured text into numerical feature vectors. The RIASEC dataset is calculated on six vocational interest scores by summing each group of responses to the personality questionnaire included in the dataset. Likewise, the datasets for Big Five calculation calculate five of the personality traits: Openness; Conscientiousness; Extraversion; Agreeableness; Neuroticism.

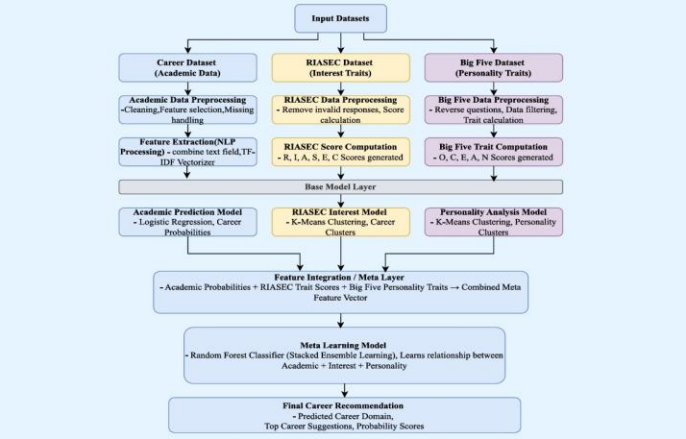

Fig. 1. *Architecture of proposed system*

Once feature extraction has been completed, base models of machine learning will be fitted to processed feature data. An academic dataset will be evaluated using a logistic regression model that predicts the probability of a person choosing each of the career fields within the dataset. In addition to this, the RIASEC and Big Five datasets will be clustered using techniques such as k-means clustering in order to

evaluate how the data clusters with respect to vocational interest and personality type. Therefore, each model will independently capture different behavioral and academic traits that are associated with career selection.

The last stage of creating a recommendation process for careers was building a stacked ensemble learning model that will combine all of the outputs of the "base" learning models together. For example, the probabilities of potential academic careers from the learning models are combined with RIASEC interest score vectors and Big Five personality trait vectors to form a meta feature vector. This combined meta feature vector is then used by a Random Forest meta classifier to classify the resulting career recommendation. This multi-layered architecture has allowed the recommendation system to describe the career prediction process by utilizing different data sources in each layer and therefore provide more complete and customized career predictions.

Data Preprocessing

The importance of data preprocessing in machine learning analysis by preparing the raw data for use cannot be overstated. This paper looked at multiple types of datasets from a variety of sources that include both structured and unstructured forms which required multiple preprocessing methods. In the case of the career recommendation dataset, preprocessing consisted of filling in missing values, standardizing attribute names and grouping job titles into related career domains so as to reduce class imbalance. Text-based attributes (i.e., course specialization, skills, certifications, interests) were combined into one text feature. After that, all of the text-based features were converted to numeric using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization in order to identify how important a word is throughout the dataset.

In regards to the RIASEC dataset, that particular dataset was preprocessed by utilizing the responses provided from vocational interest surveys to produce six standard personality measures representing the six Holland personality dimensions; Realistic, Investigative, Artistic, Social, Enterprising and Conventional. Each personality measurement was the average of responses given to similar questions on the surveys.

With respect to the Big Five personality dataset, the preprocessing steps utilized consisted of converting negatively worded questions to positively worded questions plus generating aggregate scores for the five Big Five personality dimensions; Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. The aggregate scores give numerical values to the individuals involved in measuring their psychological characteristics. Additionally, data filtering and normalization techniques are applied to remove inconsistent responses and ensure data quality across all datasets.

Base Model Development

- Academic Prediction Model: The academic prediction model analyzes data about previous study and skill-based data to make recommendations regarding potential career choices. For the textual features, they have been converted to feature vectors with a TF-IDF method. For the numerically-based features (such as the person's grades), they are included as part of the feature set. The base prediction model is a Logistic Regression classifier because it is effective for multi-class classification tasks and provides output in the form of probabilities.
- RIASEC Interest Model: The RIASEC model identifies career interests by analyzing the personality dimensions of an individual based on their personality type. Based on the calculated RIASEC traits and the use of a K-Means cluster analysis to classify individual interests into clusters based on similar patterns, these clusters represent different categories of interests and could be used to associate that type of personality to career types.
- Big Five Personality Model: The big five is a model that describes the stable psychological traits that affect behavior at work and overall career happiness. Personality vectors are used to analyze personality traits through the RIASEC model using clustering methods until personality patterns are identified. The relationships between the specific personality traits and the activities of a career are determined using the identified personality patterns.

Feature Extraction

Feature extraction plays an important role in the proposed system as it converts raw data from various data sets into structured numerical forms so that machine learning algorithms can process it. As this system integrates many different data sources such as textual academic records and answers to psychometric surveys there are different feature extraction methods to use to extract valuable patterns from these data sets. The overall goal of this stage is to convert unstructured or semi-structured data files into machine readable (interpretable) feature vectors while keeping the semantic and behavioural context of the data so that accurate career predictions can be made.

1. Textual Feature Extraction using TF–IDF

The career recommendation dataset has a number of textual attributes (academic degree, technical skills, certification, and area of interest) which all have keywords that indicate a student's level of technical competency and/or their career orientation available to be converted

into numeric form. To convert this unstructured text to a numeric format, the system uses Term Frequency - Inverse Document Frequency vectorization (TF-IDF). TF-IDF is a commonly used method in the field of Natural Language Processing (NLP). It is used to determine how critical a word is to a specific document relative to all the documents in the same corpus. In the TF-IDF measurement of a word's frequency, one component is calculated based on how frequently that word appears in a particular document and the other component is based on how often that word appears across all documents in the ETS career recommendation dataset. The weight of a word is reduced if it appears frequently throughout the entire corpus. Thus, words that contain very strong relationships within the various career domains are given more weight.

In this particular system, the multiple textual fields of skill, course specialization, certification, and area of interest are all combined or concatenated together from a student's record into a single piece of text. The TB-IDF vectorizer is then used to produce high-dimensional feature vectors that describe the importance of the various technical/components to the career domains. The resulting feature vectors can be used by increasingly sophisticated machine-learning models to relate academic performance credentials to career-type domains.

2. Integration of Numerical Academic Features

Alongside the text features in the career dataset, the data contained academic percentage scores and performance indicators that are quantitative measures of a student's academic ability and achievements. To be able to add these variables into the machine-learning model, the numerical data will have to first be normalized and then combined with the TF-IDF feature vectors.

Having both a student's qualitative academic descriptions (i.e., skills and specialization) and quantitative data (i.e., academic scores) combined in a manner as described provides the academic model with a better overall description of a student's full history with regards to education and is, therefore, more likely to enhance the accuracy of the predictions made by the academic model.

3. RIASEC Trait Feature Representation

The RIASEC dataset is used to extract vocational interest patterns based on Holland's personality theory. Each record in the dataset contains responses to a series of personality assessment questions designed to measure six vocational interest dimensions:

- Realistic
- Investigative
- Artistic
- Social
- Enterprising
- Conventional

Feature extraction for the RIASEC dataset involves computing aggregated scores for each of these six traits. This is achieved by averaging the responses associated with questions belonging to each personality category. This feature vector is composed of six dimensions, which reflect the dominant vocational interests of individuals. These features offer significant insight into various aspects of work environments or activities that individuals would likely prefer. For example, individuals with high scores related to "Investigative" features may have a stronger affinity towards research-oriented professions, while individuals with high "Enterprising" feature scores may have a stronger affinity towards leadership-oriented professions.

4. Big Five Personality Feature Representation

The Big Five personality dataset is a representation of stable personality traits using the OCEAN model. Similar to the RIASEC dataset, aggregated raw responses from personality questionnaires are used to calculate trait scores for five personality dimensions:

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

During the feature extraction process, the responses given for the negatively phrased questions are changed so that consistency is maintained in the final scores. Once the normalization process is complete, the final scores are obtained by averaging the responses for the particular traits. The final scores are a five-dimensional feature vector representing the behavioral aspects of the individual. The Big Five features are used in the system so that the factors affecting the career satisfaction are taken into account.

5. Combined Feature Representation

After extracting features from each dataset, the resulting vectors are integrated into a unified feature representation. The final feature set therefore consists of three major components:

- Academic Feature Vector (TF–IDF + numerical attributes)
- RIASEC Vocational Interest Vector (6 personality traits)
- Big Five Personality Vector (5 personality traits)

Overall, the representation of the person is based on a combination of different characteristics; they include educational ability, type of job they prefer to do and what type of personality they have. These characteristics are combined with other features into an integrated characteristics space that serves as input for the machine learning algorithms used for predicting what jobs people will have based on all of the above. The ability to take into account multiple characteristics at the same time allows this system to model complex relationships among all of the aforementioned variables thereby improving both prediction accuracy and the ability to understand why a person may choose one career path over another.

Model Development

When data has been extracted, machine learning models will be produced to classify and analyze suitability for careers by utilizing the characteristics of tools, vocational interests, and personality traits. Since this tool has used multiple different databases, each being different but still forming part of the overall project, a separate model will be created for processing each database independently. Once these models have learned how to identify features defined by their own datasets, they will work in tandem to build a finished prediction.

1. Academic Career Prediction Model

An Academic Dataset is created to develop a supervised classification scheme for predicting prospective career areas using student academic experience and skills. While other strategies could be considered for classification, we prefer the use of Logistic Regression as our main classification technique because of its ability to quickly arrive at a conclusion, ease of interpretation, and overall effectiveness in multi-class classification scenarios.

Logistic Regression predicts the likelihood that an instance belongs to a career category, given a feature value. The prediction made by the Logistic Regression model will predict the probability that a student with certain academic characteristics will fall into a particular career category.

The input features for the model will consist of both TF-IDF representations of all relevant text and numerical performance indicators for each student, as well as any additional numeric attributes (e.g., age, year graduated) that could be useful in predicting the student's career classification.

Training will consist of building the model using the labeled examples in the Career Data set where results are known. The output of the model will yield a probability distribution over multiple career categories so the system can determine a student's most likely career categories.

The advantages of using Logistic Regression in this context include:

- Ability to handle high-dimensional TF–IDF feature spaces
- Efficient training even on large datasets
- Probabilistic outputs that can be integrated into ensemble models
- High interpretability of feature importance

2. RIASEC Vocational Interest Model

The RIASEC dataset helps understand interests related to careers through Holland's vocational personality theory. The dataset focuses on measuring the personality traits of applicants and not providing them with any specific career labels; therefore, an unsupervised method was used to recognize any patterns from the information in the dataset. In this case, K-Means clustering was utilized as part of the proposed system. This is a common clustering technique that provides groupings based on similarity from six-dimensional vectors from the data set. Each data point (i.e., participant in the data set) is assigned to a group based on how similar their vocational interests are.

The clustering process involves the following steps:

- Selecting an appropriate number of clusters k.
- Initializing cluster centroids randomly.
- Assigning each data point to the nearest centroid based on distance metrics.
- Updating centroids iteratively until convergence is achieved.

The clusters that are formed represent different vocational interest groups. For example, a cluster with high scores for Investigative and Realistic may represent people with a vocational interest in technical or analytical work. These clusters provide useful insights into the

types of work associated with certain interest groups.

3. Big Five Personality Analysis Model

The behaviors that one tends to have as a person will define the kind of jobs one will prefer and how one will be in the future at the workplace, and this is referred to as the Big Five Personality Dataset. This is very similar to the RIASEC Dataset because it provides one with information about their personality and not a career. Clustering algorithms have thus been used to establish the significance of the personality traits. The K-means clustering algorithm has thus been used to cluster the 5-Dimensional Personality Feature Vectors, which include Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Every cluster reflects a unique personality type, which may have implications for particular career environments.

- High Openness & Conscientiousness may fit well with research/analytical-type professions.
- High Extraversion & Enterprising traits may fit well with leadership/management-type professions.

Personality Clusters offer a further dimension of behavioral characteristics, which can be related to academic & vocational interest features.

4. Ensemble Career Prediction Framework

The final career recommendation is derived from the final stage of the methodology, which combines all the results from all of the different individual models. The reason for combining these results is that the individual models' datasets can each capture a separate and distinct aspect of a person's suitability to a career. The method used to combine the results of all of the individual models will use a technique called stacked ensemble learning.

5. Meta-Feature Construction

Meta-features are a consolidated feature representation made from combining models base outputs. The components of the meta-features include:

- The academic prediction model's generated career probability scores.
- RIASEC interests scores.
- Big Five Personality trait scores.

Thus, the meta-feature produces a multidimensional representation that combines academic skills with vocational interests and personality traits.

6. Meta-Learning Model

Once combined, these meta-features will be used to train a Random Forest classifier which will act as the ensemble framework's meta-model. Random Forest is an ensemble learning algorithm that creates multiple decision trees that combine together to provide more accurate predictions. The combination of the decision trees will be created from a random subset of both the data used and the features using the data in order to minimize overfitting and improve accuracy.

The key advantages of using Random Forest in this stage include:

- Ability to deal with high dimensional feature spaces
- Resistance to noise and overfitting
- Ability to represent complex non-linear relationships
- Increased performance through ensemble methods during prediction purposes

A Random Forest model has learned how particular types of academic attributes, personality traits, and individual preference when selecting a vocation impact an individual's outcome.

7. Final Career Recommendation

Once trained, the meta-model can make predictions about new users. When a new user enters academic information and completes personality assessments, the following steps take place:

- The academic model extracts academic features and generates the probabilities of possible career paths based on those features.
- The model calculates each user's RIASEC vocational interest scores.
- The model calculates each user's Big Five personality trait scores.
- All of the calculated values are combined to form a meta-feature vector.
- The Random Forest meta-model receives the meta-feature vector so it can make a final career path recommendation for the

user, along with probability scores associated with how likely that user would be to find success in specific fields.

The final recommendation is made using an ensemble-based approach, meaning that multiple independent evaluations of the meta-feature vector are made before the final recommendation is provided.

Results

Overall Model Performance

To evaluate how well the CareerLens v3 stacked ensemble model predicts, how stable it is, and how well it performs, statistical metrics were used. At least two forms of validation were used for the model's validation (5-fold and 10-fold), and there was also a separate dataset used for testing to confirm that the results are reliable and not biased.

The CareerLens v3 stacked ensemble model received an accuracy rating of 94.68% (±0.0049) for the five (5) fold validation and an accuracy rating of 94.78% (±0.0081) for the ten (10) fold validation. The accuracy results are similar, suggesting that the CareerLens v3 stacked ensemble model has proven to be stable in regard to all datasets. Moreover, the model received an accuracy of 95.51% on the test dataset, supporting the claim that it can generalize well among multiple datasets.

Additional metrics were also assessed to further evaluate the performance of the CareerLens v3 stacked ensemble model. The model received a precision rate of 0.9553, a weighted recall of 0.9551 and a weighted F1-score of 0.9551. These three additional metrics demonstrate that the CareerLens v3 stacked ensemble model has provided a good predictive solution and performance against all classes. Finally, the macro F1-score of 0.9562 provides further proof that the CareerLens v3 stacked ensemble model has consistently performed well among all career domains.

Validation of the CareerLens v3 model is supported by satisfactory performance measurements: Cohen Kappa = 0.9501 and Matthews Correlation Coefficient = 0.9501. Both measures indicate strong reliability for the CareerLens v3 model. Based on this model's high macro ROC-AUC value of 0.9978, we conclude that the CareerLens v3 model does a very good job of distinguishing between the different career domains in question. Overall, these results imply that the CareerLens v3 stacked ensemble model has excellent predictive power and generalizability.

Per-Class Performance Analysis

There were individual evaluations conducted across each of the ten fields of work. The assessment revealed that the CareerLens v3 model is able to produce similar results across each classification type. In the creative arts and design area's assessment scores produce an F1 score equal to 0.99. Type classifications for financial and accounting, health care and medical, and marketing and sales all displayed strong classification scoring results as well.

In contrast, the model provided lower scoring results for the data science and artificial intelligence (F1=0.90) and IT Software (F1=0.92) categories. This could be attributed to the close degree to which the personality traits exhibit similarities in each of those industries. Despite these poor classifications CareerLens still maintains a well-balanced score.

Table 1. *Per-Class Evaluation Metrics for Career Domain Classification*

Career Domain	Precision	Recall	F1-Score	Support
Business Management	0.94	0.97	0.95	172
Core Engineering	0.97	0.95	0.96	178
Creative Design & Arts	0.98	1.00	0.99	164
Data Science & AI	0.90	0.90	0.90	181
Education & Research	0.96	0.96	0.96	169
Finance & Accounting	0.99	0.97	0.98	164
Healthcare & Medicine	0.97	0.97	0.97	162
IT Software	0.91	0.93	0.92	195
Law & Public Service	0.96	0.95	0.95	162

Confusion Matrix Analysis

A confusion matrix provides visual representation of how well a classifier is performing by comparing predicted values to actual (true) values for each class of data. As can be seen in the confusion matrix below, there is quite a lot of data located along the diagonal (the area

where vertical and horizontal lines intersect), indicating that this classifier is producing the majority of its classifications correctly. In fact, each class produced by this classifier has almost no misclassifications. However, two classes produce small amounts of misclassifications:

- Data Science & Artificial Intelligence and Information Technology.
- Business Management and Marketing and Sales.

These types of errors are to be expected due to similarities between the characteristics of the various classes, as well as because of the similarities/matching domains of various classes. For example, both people who work with Data Science/Artificial Intelligence and Information Technology tend to have similar skillsets (e.g. analytical). Similarly, individuals working in Business Management and those working in Marketing and Sales both exhibit many of the same characteristics (e.g. enterprising personality type/social personality type of individuals).

Fig. 2. *Confusion Matrix Illustrating Per-Class Prediction Performance of the Stacked Ensemble Model*

Baseline Model Comparison

In order to evaluate its performance, we have compared the newly developed model with its competitors of similar types: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbor, and Naïve Bayes.

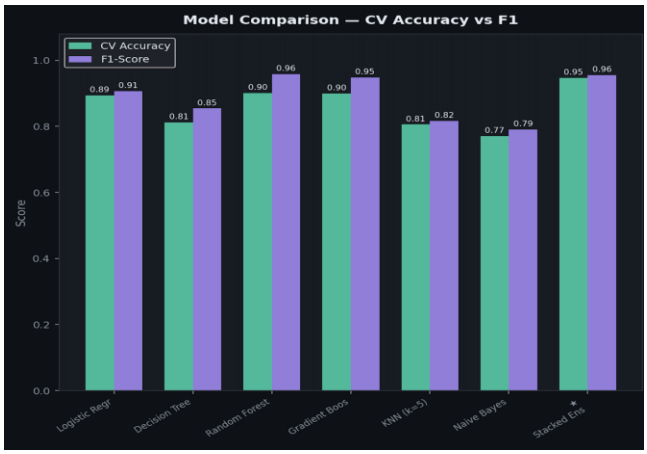

Fig. 3. *Comparative Analysis of ML Models on Accuracy and F1 Metrics*

Based on the comparative results, the following conclusions can be drawn:

1. The model that was stacked achieved an accuracy of 94.68% through cross-validation; therefore it can be said that it is superior to other models.
2. The Random Forest model achieved slightly higher accuracy (95.74%); however this model does not have the stability of the stacked model.
3. The benchmark models (e.g. Naive Bayes, KNN) lack the ability to effectively handle the complexity of relationships present in the data and thus are far less accurate than the stacked model.

The accompanying graphic clearly demonstrates that the stacked model provides the best compromise between accuracy and stability therefore it should be considered the most reliable model.

Ablation Study

An ablation study was conducted to evaluate the contribution of different feature groups: Academic data, RIASEC traits, and Big Five personality traits.

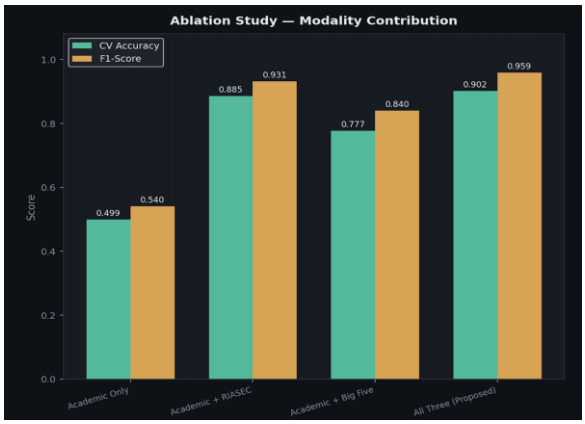

Fig. 4. Ablation Study on Modality Contribution Using CV Accuracy and F1-Score

The results show that:

- Using academic data alone results in poor performance (Accuracy ≈ 54%), indicating that academic scores are insufficient for career prediction.
- Adding RIASEC features significantly improves performance (≈ 93%), demonstrating that interest-based profiling is a critical factor.
- Adding Big Five traits provides moderate improvement, but less than RIASEC.
- The combined model (Academic + RIASEC + Big Five) achieves the highest performance (95.86% accuracy).

Figure 4 visually confirms that multi-modal integration leads to substantial performance improvement, validating the design of the proposed system.

Feature Importance Analysis

The importance of each feature group in relation to how much they contributed to the final predictions was analysed and the following results were obtained:

- RIASEC features = 44.2%
- Academic features = 40.7%
- Big Five traits = 15.2%

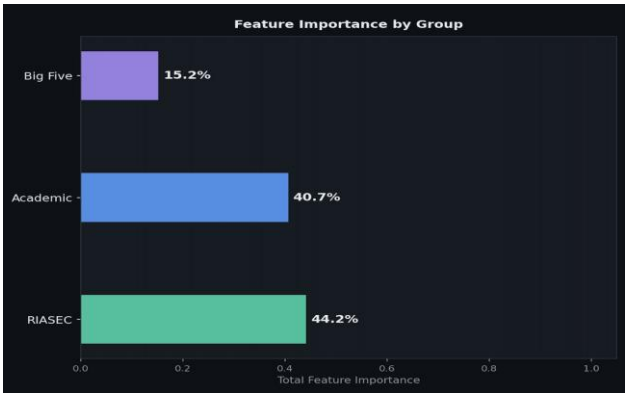

Fig. 5. *Comparative Importance of Academic, RIASEC, and Personality Features*

Figure 5 illustrates that RIASEC features are by far the most impactful, demonstrating that interest-based alignment with careers is important. Academic performance also contributes significantly while personality traits serve to assist with refining the prediction. Additionally, Figure 5 creates a visual indication of this distribution and represents that the RIASEC features are the clearly dominant feature group.

System Output and Final Conclusion

Developed CareerLens v3 creates tailored career projections based on a combination of an individual's academic achievements, RIASEC interests, and Big Five traits. In the table below, Core Engineering (27.27%), Data Sciences and AI (12.2%) and Business and Management (8.12%) show the three most appropriate career domains based on the given user's profile as indicated by the degree of confidence associated with each of those careers. The CareerLens v3 system also gives insight as to how much each element of input contributed to the ultimate decision, thus providing users with confidence in the system's ability to accurately predict user-projected careers; the RIASEC interest contributed.

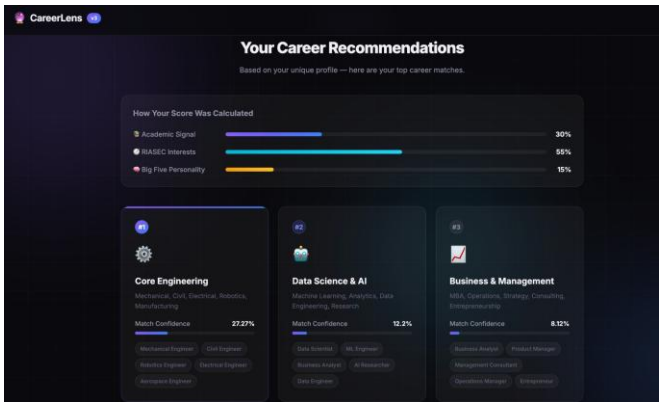

Fig. 6. *Output Visualization of the Proposed Career Guidance System*

55%towards the overall prediction, academic attributes contributed 30% and personality traits contributed 15% towards the overall career prediction.

In summary, the findings indicate that the newly designed stacked ensemble approach offers both accurate predictions (~95%) as well as valuable career recommendations that are explainable and focus on the end user. Multi-faceted features enhance the quality of decisions made compared to previous unidimensional data sources. Therefore, the system has been shown to be an efficient, scalable, and practical source of providing intelligent support for students making educated and individualistic career choices.

Conclusion

The proposed Self-Improving AI-Based Career Suggestions System would benefit from Quality Education and Effective Governance through intelligent, personalized, and data-driven career suggestions to assist students. With the ability to deliver suggestions based on an analysis of students' skills, interests, personality and academic history, this system reduces reliance on traditional manual and biased methods of counselling. It would also assist schools and higher education institutions with improved scalable and adaptive career guidance mechanisms.

In addition, the proposed system would benefit from a governance perspective through identifying trends of employability, skill shortages, and emerging career requirements. A self-improving learning mechanism would improve recommendations by taking into account new data and user feedback. The proposed system would reduce career misalignment and improve employability of all ages. Societally, the proposed system would assist in achieving Sustainable Development Goals such as Quality Education, Decent Work, Reduced Inequality and is an effective scalable solution for today's learning and career guidance systems.

Future scope

- Integration with Professional Platforms: The future versions of that system have the ability to incorporate professional networking platforms like LinkedIn, as well as job portals, to gather real-time job-related data, professional profiles and industry-related requirements for career recommendations which will be more accurate.
- Adoption of Deep Learning Models: With regard to incorporating more advanced forms of deep learning techniques like neural networks and transformer-based natural language processing models; each of which will enable the system to achieve greater semantic analysis accuracy with respect to resumes, skills, and career description for increasing the accuracy of predictions.
- Skill Gap Analysis and Personalized Learning Paths: Future additions could include modules that will enable the identification of the skill gaps that exist between the user's current experience and their target role, and recommend appropriate training courses, certificates or other learning resources.
- Real-Time Industry Data Integration: The system can be enhanced to include integration of real-time data directly from online job boards and corporate industry data sources for increasing the accuracy of job recommendations in terms of current industry

demands and new job positions.

- Scalable Web and Mobile Deployment: Because this system is generally available as a cloud based application, its scalability will allow the greatest usage of the system by a large population of students, all universities and all career related counseling applications.

References

1. S. H. Faruque, S. A. Khushbu, and S. Akter, "Unlocking futures: A natural language driven career prediction system," *arXiv preprint arXiv:2405.18139*, 2024.
2. R. Guru, M. Devi, R. N. Srivastava, D. Salwan, and A. Agarwal, "AI-driven career guidance system using psychometric profiling and machine learning," *Int. J. Sci., Eng. Technol.*, 2025.
3. A. Shahzada *et al.*, "CAREERLLAMA: An AI-powered personalized career recommendation system," *Spectrum of Engineering Sciences*, vol. 3, no. 6, pp. 404–414, 2025.
4. K. S. V. Reddy *et al.*, "Resume analyzer and job recommendation system," *Iconic Research and Engineering Journals*, vol. 8, no. 9, pp. 282–287, 2025.
5. R. Sandra *et al.*, "Smart career advisor: A machine learning based recommendation system," *Int. J. Sci. Res. Sci. Technol.*, vol. 12, no. 15, pp. 328–337, 2025.
6. G. Manikandan, V. Veronica, and S. Hemalatha, "Integrating learning analytics and recommendation models for career guidance," *Int. J. Sci. Res. Sci. Technol.*, vol. 11, no. 2, pp. 169–176, 2024.
7. P. Shenoy, R. Shiroorkar, and A. Samani, "CareerLink: AI for resumes and career advice," *IJRASET*, 2025.
8. A. Kulkarni *et al.*, "Resume-based job recommendation system using NLP and deep learning," *IJRASET*, 2024.
9. P. Singh Sisodiya *et al.*, "AI-powered personalized career guidance and skill recommendation system," *AIJMR*, vol. 3, no. 6, 2025.
10. Z. Wang *et al.*, "CareerPooler: AI-powered career exploration system," *arXiv preprint arXiv:2509.11461*, 2025.
11. X. Q. Ong and K. H. Lim, "SkillRec: A data-driven approach to job skill recommendation," *arXiv preprint arXiv:2302.09938*, 2023.
12. "AI-powered career guidance system using machine learning," *IJIREM*, 2025.
13. "AI-driven career guidance system for student recommendations," *Frontiers in Health Informatics*, vol. 13, no. 3, 2024.
14. "AI-powered career recommendation for students: A survey," *IJCSE*, vol. 9, no. 6, 2025.
15. T. Yadalam *et al.*, "Job recommendation systems through AI and machine learning," *Int. J. Basic Appl. Sci.*, 2025.
16. "An AI-driven approach for personalized major and career recommendation," *ACM*, 2025.
17. "Machine learning for educational and career guidance: A review," *Frontiers in Education*, 2025.
18. D. F. Mujtaba and N. R. Mahapatra, "Fairness in AI-driven recruitment," *arXiv preprint arXiv:2405.19699*, 2024.
19. S. A. Alsaif *et al.*, "Learning-based job recommendation systems," *Computers*, vol. 11, no. 11, 2022