AI-Based Career Guidance System

Jatin, Subesh Upadhyay, Aditya Raj Dixit  
Lovely Professional University, Punjab  
Department of Computer Science and Engineering  
March 2025

# Abstract

This research presents the comprehensive design and implementation of an AI-powered Career Guidance System that aims to simplify and personalize one of life’s most critical decisions—choosing the right career. Recognizing the limitations of traditional counselling methods, which often rely on generalized assessments or outdated psychometric techniques, our system integrates machine learning with an intuitive web interface to offer a smarter, more tailored approach.

By allowing users to self-evaluate on a wide range of both technical skills (like programming, communication, and project management) and soft personality traits (such as openness, agreeableness, and emotional range), the system constructs a multidimensional profile of the individual. This profile is then analysed by a machine learning model—specifically, an XG Boost classifier—trained on an extensively curated and balanced dataset that includes 26 distinct career paths. The model not only considers the technical fit but also psychological alignment, ensuring that the recommendations go beyond what a resume or test score can tell.

With a remarkable classification accuracy of 98.28%, the model significantly outperforms other popular algorithms including Random Forest, KNN, and SVM in both precision and generalization. But beyond just numbers, the system shines in its user experience: built using Stream lit, the web interface is clean, interactive, and engaging. Users can adjust sliders to explore different skill and personality profiles, and instantly see the top three career roles that best match their inputs.

This solution bridges the often-overlooked gap between aptitude assessments and real-world career guidance, making it a valuable tool for students, educators, institutions, and professionals alike. Ultimately, it empowers individuals not just with data—but with direction.

# Keywords

Career Guidance, Machine Learning, XGBoost, Streamlit, Personality Traits, Skills Assessment

# 1. Introduction

The process of selecting a career path has long been a challenge for students and professionals alike. Traditional career counseling techniques often lack the personalized touch and fail to incorporate real-time market demands. With the emergence of artificial intelligence and machine learning, it has become feasible to automate and personalize career recommendations. This research introduces an intelligent career guidance model that uses technical skills and psychological traits to predict optimal career paths. It leverages the XGBoost classifier for high accuracy and is integrated into a user-friendly interface using Streamlit. The combination of predictive power and interactive UX makes the system highly practical and accessible.

Notable recent works in the career recommendation domain include systems that integrate collaborative filtering, deep learning, and hybrid recommendation techniques. However, many of these approaches lack explainability or fail to provide real-time interactivity. In contrast, our system provides transparency and instant feedback through an interpretable model and a frontend interface.

# 2. Related Work

Several approaches have been proposed for automated career guidance. Early methods used rule-based systems and statistical heuristics to make recommendations. These systems were limited by static rules and an inability to adapt to nuanced personality and skill profiles. More recently, researchers have employed machine learning models such as decision trees, logistic regression, and clustering methods for matching users to career paths. However, these models often underperform due to inadequate feature engineering or lack of personalization. Our work builds upon these approaches by integrating a larger feature space—including soft skills and personality traits—into a high-performing model trained using XGBoost.

The methodology is structured into three core modules: data processing, model training, and frontend integration. The dataset is extended using synthetic generation for underrepresented career roles to balance class distribution. Features were selected based on expert consultation and correlation analysis. Data normalization and label encoding ensure uniformity and compatibility with XGBoost, which handles multiclass classification efficiently.

# 3. Proposed Methodology

The development of this AI-based Career Guidance System followed a carefully structured methodology that blends data science with user-centred design. The process begins with the construction of a robust dataset that captures both technical and behavioural attributes relevant to real-world career success. Each entry in the dataset represents a hypothetical or real individual—be it a student or a working professional—and includes feature scores such as Programming Skills, Communication Abilities, Project Management expertise, and a range of personality dimensions like Openness, Emotional Range, Extraversion, and Self-Transcendence.

To prepare the dataset for modelling, a series of preprocessing steps were implemented. All numeric features were standardized using Standard Scaler to bring them onto a consistent scale, ensuring that no single feature would dominate the learning process. The target labels—career roles—were encoded into numerical format using Label Encoder, allowing them to be processed by machine learning algorithms.

Once the dataset was cleaned and structured, multiple classification models were evaluated. These included Logistic Regression, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbours (KNN). After rigorous testing and cross-validation, the XG Boost classifier stood out as the best performer, achieving an impressive test accuracy of 98.28%. Its ability to handle multiclass classification, combined with regularization and boosting techniques, made it exceptionally suited for this task.

To bring the model to life in a way that users could engage with meaningfully, it was embedded into a Stream lit web application. The interface is designed for simplicity and ease of use. Users are invited to rate themselves on a series of sliders that represent both soft skills and technical proficiencies. Once submitted, the system instantly scales and transforms the inputs, passes them through the trained XG Boost model, and returns the most likely career role. Additionally, to support broader decision-making, the interface also presents the top three career recommendations, ranked by confidence scores.

This methodology ensures that the system is not only technically robust but also highly usable, making personalized career guidance more accessible, interactive, and impactful than ever before.

# 4. Implementation & Tools

The system was developed using **Python**, leveraging its versatility and rich ecosystem of data science libraries. The machine learning model and preprocessing pipeline were built using **Scikit-learn**, **XG Boost**, **NumPy**, and **Pandas**, while the model was serialized using **Job lib** for efficient deployment.

The **frontend** was created using **Streamlit**, chosen for its simplicity and ability to convert Python scripts into fully interactive web applications. The UI features a modern two-column layout with **sliders** for 14 input features, allowing users to easily rate themselves on both technical competencies and personality traits. These inputs are then passed through a preprocessing layer before being fed into the trained model.

The **machine learning architecture** consists of an **XGBoost classifier**, which was trained on a carefully curated and extended dataset. The dataset was balanced using synthetic data generation techniques to ensure fair representation across all 26 career roles. XGBoost was chosen after benchmarking against other models such as Random Forest, Logistic Regression, SVM, and KNN due to its superior performance in terms of both accuracy and generalization.

Deployment was initially conducted on a **local system**, but the entire system is lightweight and portable, making it ideal for hosting on platforms like **Streamlit Cloud** for broader accessibility.

**Performance Evaluation**

To ensure that the model was not just accurate but also reliable and fair, a range of **evaluation metrics** were used:

* **Accuracy:** The model achieved a stellar accuracy of **98.28%**, correctly classifying the vast majority of test samples.
* **Precision, Recall, and F1-Score:** These metrics were computed for each class. The model consistently achieved **F1-scores above 97%**, even for minority roles.
* **Confusion Matrix:** Detailed analysis showed minimal confusion between similar roles, indicating that the model could distinguish even subtle differences in user profiles.
* **Cross-Validation:** A **5-fold cross-validation** strategy was applied to assess model stability. Accuracy across all folds remained consistently above 97%, affirming that the model is robust and not overfitted.

Together, these implementation and evaluation strategies ensure that the system is not only technically sound but also practical, fair, and scalable for real-world use.

# 5. Results & Discussion

The final model, trained using the XGBoost classifier, achieved an impressive **accuracy of 98.28%** on the test set, demonstrating a clear advantage over other machine learning algorithms. When benchmarked against **Random Forest (97.6%)**, **Support Vector Machine (94%)**, and **K-Nearest Neighbors (94.9%)**, XGBoost consistently outperformed in terms of precision, recall, and overall generalization. This superior performance is attributed not only to the strength of the algorithm but also to thoughtful **feature selection**, **data balancing**, and **engineering strategies** such as SMOTE-based augmentation.

Despite the initial class imbalance in the dataset, the system showed exceptional generalization across all 26 career roles. This was validated through cross-validation and confusion matrix analysis, which revealed strong classification performance even on minority classes. Such robustness is critical when deploying AI solutions that cater to diverse user profiles.

In real-world simulations, the **Streamlit web application** emerged as a strong complement to the backend model. It does more than just return a single prediction—it offers users a **ranked list of the top 3 most compatible careers**, each with confidence scores. This design not only enhances transparency but also gives users room to explore multiple paths based on their unique combination of skills and personality. The interface was found to be intuitive, responsive, and insightful during user testing, making it accessible even to non-technical audiences such as students and counselors.

# 6. Conclusion

This project clearly demonstrates the transformative potential of combining **artificial intelligence** with **human self-assessment** for career guidance. By training a machine learning model on a dataset enriched with both technical skills and psychological traits, we have created a system that surpasses the rigidity and limitations of traditional counselling approaches. The result is a digital career coach that doesn’t just provide answers—it empowers individuals to better understand themselves and explore options aligned with their strengths.

The real-time, user-friendly deployment through Streamlit enhances its accessibility, making it a practical tool for institutions, educators, and career consultants. Whether it's used in classrooms, counselling offices, or job readiness programs, the solution is ready to scale and evolve.

In a world where careers are dynamic and individual potential is multifaceted, our system offers not just predictions, but possibilities—and that’s the future of career guidance.

# 7. Future Work

While the current system provides a powerful foundation for intelligent career guidance, there are several exciting directions in which this project can evolve to become even more impactful, inclusive, and intelligent.

1. **Integration with Real-Time Job Market APIs:**  
   One of the most promising enhancements involves connecting the system to live job market data through APIs such as LinkedIn, Glassdoor, or Indeed. This would allow the model to not only suggest careers based on a user’s profile, but also weigh recommendations based on **current industry demand**, regional hiring trends, and future job outlook. Imagine a system that tells a user not just what they’re good at—but also what’s in high demand today and tomorrow.
2. **Expanding Dataset to Include Multilingual and Multicultural Inputs:**  
   To truly serve a global audience, it’s important that the system reflects the diversity of users across cultures, languages, and educational backgrounds. Future versions of the dataset will include responses and ratings from people across different geographies, enabling the model to provide **culturally sensitive and localized recommendations**. This expansion will also help break language barriers and make the platform more inclusive.
3. **Incorporating Deep Learning Models:**  
   While XGBoost has performed exceptionally well, exploring deep learning architectures—such as neural networks, attention-based models, or even transformers—could open up new possibilities. These models are better at handling **nonlinear relationships and higher-dimensional data**, potentially boosting prediction accuracy and uncovering hidden patterns in user traits.
4. **Cloud Deployment and Platform Integration:**  
   The ultimate goal is to make this tool available to as many users as possible. Hosting it on platforms like **Streamlit Cloud, AWS, or Azure** would ensure accessibility and scalability. Moreover, **integrating it with career ecosystems** like LinkedIn or academic portals could allow for real-time syncing of user profiles, resume data, and job preferences—creating a seamless experience from guidance to job discovery.

# 8. References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, 'An Introduction to Statistical Learning,' Springer, 2013.  
[2] L. Breiman, 'Random Forests,' Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.  
[3] XGBoost Documentation: https://xgboost.readthedocs.io/  
[4] Streamlit Documentation: https://docs.streamlit.io/  
[5] Scikit-learn Documentation: https://scikit-learn.org/

# 9. System Architecture

The AI-Based Career Guidance System has been architected with a **modular, scalable, and future-ready design**, enabling seamless integration, flexible expansion, and cloud deployment. It follows a **three-tier layered architecture** that clearly separates concerns and responsibilities, making the system maintainable and robust.

**1. Data Layer – The Intelligence Core**

At the heart of the system lies the **data layer**, which houses a rich and diverse dataset composed of both **technical skill ratings** (e.g., Programming Skills, Project Management) and **personality traits** (e.g., Openness, Agreeableness, Emotional Range). Each entry in this dataset reflects a unique individual profile, paired with a recommended career role. This layer serves as the foundational knowledge base that fuels the machine learning model.

**2. Machine Learning Processing Layer – The Analytical Brain**

The ML processing layer is where the real intelligence unfolds. This layer handles:

* **Preprocessing tasks**, such as feature normalization using StandardScaler and categorical encoding with LabelEncoder.
* **Model training**, where multiple algorithms were evaluated and XGBoost was selected for its superior performance.
* **Inference logic**, which enables the model to make predictions based on new inputs.
* **Evaluation mechanisms**, including accuracy, F1-score, confusion matrix, and cross-validation, to ensure model reliability.

This layer acts as the system’s analytical engine, continuously learning patterns and mapping user profiles to the most suitable career outcomes.

**3. Presentation Layer – The User Interface**

The final layer is the **presentation layer**, built using **Streamlit**, which transforms the backend intelligence into an **interactive, user-friendly web experience**. Here, users can rate themselves via sliders representing each feature, and upon submission, the model processes the input in real time to display:

* The most suitable career role
* A confidence score
* A ranked list of the top 3 career matches

The simplicity and responsiveness of this layer ensure accessibility for users across all backgrounds, from high school students to academic advisors.

**Scalability & Integration Capabilities**

The system is designed with **loose coupling**, ensuring that each layer can be modified or upgraded independently. For example:

* The ML model can be retrained without altering the UI.
* The UI can be redesigned without touching the backend logic.

Moreover, the architecture is **cloud-ready**. It supports deployment on platforms like **Streamlit Cloud, AWS, or Azure**, and can be extended with **APIs** for integration with external services such as LinkedIn, job boards, or academic portals.

This design future-proofs the system, making it adaptable to evolving technologies and user needs.

# 10. Dataset Description

At the core of this AI-based career recommendation system lies a thoughtfully constructed dataset that captures the diversity and depth of human potential. Initially, the dataset was built using **16 carefully selected features**, representing a balance of **technical skills** (e.g., Programming Skills, Project Management, Computer Architecture) and **behavioural attributes** (e.g., Openness, Agreeableness, Emotional Range). These features were chosen to reflect real-world criteria that influence professional success, beyond just academic scores or degrees.

Each data instance represents a student or professional profile, labelled with a corresponding **career role**. The original version of the dataset mapped to **16 common career paths**, including roles such as Software Developer, Business Analyst, and Cybersecurity Specialist.

To make the system more relevant to today’s fast-evolving job market and to enhance the **generalizability** of the model, we **expanded the dataset by introducing 10 new trending career roles**. These include emerging fields like:

* Prompt Engineer
* Cloud Solutions Architect
* Game Developer
* NLP Specialist
* Digital Marketing Strategist
* UI/UX Designer, among others.

This brought the total number of unique career classes to **26**, covering a much broader spectrum of professional pathways across both technical and creative industries.

To support each role with sufficient training data, we **generated synthetic rows** using statistical sampling and controlled randomness. These synthetic instances were modelled on existing class distributions, preserving feature integrity while expanding variety. Each role was assigned a **minimum of 100 samples**, ensuring that the model could learn patterns effectively for both majority and minority classes.

To further improve data balance and fairness, **SMOTE (Synthetic Minority Over-sampling Technique)** and manual **up sampling** were employed. This ensured that the machine learning model didn’t become biased toward well-represented roles, maintaining equitable performance across all career categories.

This enriched and balanced dataset forms the foundation of a reliable, inclusive, and adaptive career recommendation engine—capable of guiding users not just toward common roles, but also toward new, in-demand professions that align with their unique strengths and interests.

# 11. Feature Selection and Engineering

Feature engineering served as one of the **cornerstones of this project**, acting as the bridge between raw user input and meaningful career predictions. The strength of any machine learning model lies not just in its algorithm, but in how well the data represents the problem space—and in this project, **careful feature selection and transformation were key to achieving high performance**.

The dataset was designed to incorporate a holistic view of the individual by combining both **hard skills** and **soft traits**:

* **Technical proficiencies** like *Programming Skills*, *Computer Architecture*, *Project Management*, and *Communication Skills* reflect practical capabilities that directly influence job performance.
* **Behavioural and psychological dimensions** such as *Openness*, *Agreeableness*, *Hedonism*, *Emotional Range*, and *Self-Transcendence* add depth to the profile, helping to match users with roles that not only suit their skillsets but also align with their personality.

To validate the **relevance and importance of each feature**, we employed:

* **Correlation heatmaps** to identify feature dependencies and redundancies.
* **Consultations with academic mentors and career counsel lors** to ensure the psychological and professional features chosen were grounded in real-world decision-making.

Once finalized, the features were prepared for model training:

* **Normalization using Standard Scaler** was applied to ensure that features with larger numerical ranges did not dominate those with smaller scales. This is especially important in models like XGBoost where input ranges can affect learning dynamics.
* **Label Encoding** was used to convert categorical output labels (career roles) into numeric class indices, making them suitable for **multiclass classification**.

The final feature vector consisted of **14 dimensions**, each representing a unique aspect of a person’s profile. This structure allowed the model to understand users in a nuanced and balanced way—enabling not just accurate predictions, but predictions that make intuitive sense to both the user and the counsellor.

By engineering features that reflect both competence and character, the system was built to respect the complexity of real human career journeys.

# 12. Evaluation Metrics

A robust machine learning model is not defined by accuracy alone. To ensure the reliability, fairness, and real-world applicability of our AI-Based Career Guidance System, we adopted a **comprehensive evaluation framework** that goes beyond basic performance indicators. Here’s how we validated our model’s effectiveness:

* **Accuracy – Overall Prediction Success:**  
  The model achieved a remarkable **accuracy of 98.28%**, indicating that it correctly predicted the career role for the vast majority of test samples. While impressive, accuracy alone doesn’t always tell the full story—especially in multiclass settings with class imbalances.
* **Precision & Recall – Class-Specific Reliability:**  
  To dive deeper, we evaluated the model’s **precision** and **recall** for each of the 26 career roles. These metrics are crucial in identifying how well the model avoids false positives and false negatives. Impressively, **precision remained above 95% across all roles**, demonstrating that when the model made a prediction, it was very likely to be correct.
* **F1-Score – Balanced Performance Metric:**  
  The **F1-score**, a harmonic mean of precision and recall, was used to strike a balance between both metrics. The model maintained an **average F1-score of 97.5%**, further confirming its balanced performance and reliability, especially on minority classes that are typically harder to classify.
* **Cross-Validation – Ensuring Generalization:**  
  To avoid overfitting and assess how well the model would perform on unseen data, we implemented **5-fold cross-validation**. The model consistently scored above 97% in every fold, reinforcing its **stability and robustness** across different subsets of data.
* **Confusion Matrix – Insight into Misclassifications:**  
  Finally, a **confusion matrix** was analysed to visualize and understand the nature of misclassifications. This helped us identify overlapping roles (e.g., Software Developer vs. API Specialist) and fine-tune our data balancing and feature engineering strategies. The matrix showed minimal confusion even among roles with similar skill profiles.

Together, these evaluation metrics confirm that the system is not only accurate but also **consistent, fair, and capable of delivering reliable career predictions across a diverse user base**.

# 13. User Interface and Experience

The **Streamlit-based user interface** was intentionally crafted with a focus on **simplicity, accessibility, and engagement**, ensuring that users from diverse backgrounds—students, counsellors, educators, and even non-technical individuals—can interact with the system effortlessly.

At the core of the interface is a clean, form-style layout where users are guided to **self-evaluate using interactive sliders**. Each slider corresponds to one of the 14 input features, ranging from technical competencies (like Programming Skills) to behavioural traits (like Openness or Emotional Range). This design mirrors the familiar experience of filling out a digital questionnaire or self-assessment form, making it instantly approachable.

Once the user submits their inputs, the interface triggers the backend prediction model and, within seconds, displays:

* The most **suitable career role**
* A **confidence score**
* A **ranked list of the top 3 career suggestions**

This layered feedback not only boosts user trust but also allows individuals to explore **alternative paths** they might not have considered. It encourages self-reflection and open-minded decision-making, rather than presenting a single rigid outcome.

From a technical standpoint, the UI supports **real-time inference**, runs entirely in the **browser**, and **requires no software installation**, making it perfect for:

* Career guidance sessions in schools and colleges
* Human resource assessments and internal upskilling
* Integration into job readiness platforms and educational portals

Its lightweight architecture and intuitive flow also make it **ready for cloud deployment**, allowing it to scale seamlessly and reach users globally.

In short, the UI is not just a frontend—it’s an **experience layer** designed to foster confidence, clarity, and exploration in one of life’s most important decisions: choosing a career.

## Git-hub Repository

<https://github.com/jatin322-std/CareerGuidanceTool.git>