CAREER PREDICTION SYSTEM USING MACHINE LEARNING

Project Report

Submitted by:

Sreerag K

Program / Domain:

Data Science

Training Institute:

Aesthetix Edu-tech

Academic Year:

2025–2026

Contents

Abstract

Choosing a suitable career is one of the most critical decisions in a student's life. Traditional career guidance approaches rely heavily on manual counseling and subjective assessments, which may lead to biased outcomes. With the advancement of Machine Learning (ML), data-driven systems can analyze student profiles and provide intelligent career recommendations.

This project presents a Career Prediction System using Machine Learning that predicts suitable job roles based on students' skills, interests, certifications, academic attributes, and behavioral traits. Multiple supervised learning algorithms were evaluated, and the Random Forest Classifier was selected as the final model due to its superior performance and robustness. The system achieved an accuracy of 87%, outperforming the Support Vector Machine model. A pipeline-based architecture ensures consistent preprocessing and reliable predictions, enabling students to make informed career decisions.

Chapter 1

Introduction

Career selection plays a vital role in shaping an individual's professional success and personal satisfaction. In the modern job market, students are exposed to a wide range of career options, often resulting in confusion and uncertainty. Poor career choices may lead to dissatisfaction, low productivity, and limited career growth.

Traditional career guidance systems depend largely on counselors' experience and student self-assessment. These methods are subjective and do not scale well for large student populations. Machine Learning provides an effective alternative by identifying patterns in historical data and delivering objective career recommendations. This project aims to develop an automated career prediction system that assists students using machine learning techniques.

Chapter 2

Scope of the Project

- To analyze student data and identify meaningful career patterns
- To build a machine learning-based career prediction system
- To evaluate multiple classification algorithms
- To select the most accurate and reliable model
- To provide multiple career recommendations
- To design an interactive user interface using Streamlit

Chapter 3

Dataset Description

The dataset used in this project consists of structured student-related information collected to support career prediction using machine learning techniques. It contains a diverse set of attributes representing students' academic skills, technical abilities, interests, certifications, workshops, and personality traits. Numerical features include logical quotient rating, coding skills rating, number of hackathons participated in, and public speaking points, which help assess a student's analytical and communication abilities. Categorical features capture behavioral and learning characteristics such as self-learning capability, teamwork experience, introversion, reading and writing skills, and memory capability score.

In addition to structured attributes, the dataset includes text-based features such as interested subjects, certifications, workshops attended, preferred company type, interested career area, and types of books read. These features provide insight into a student's career inclinations and domain preferences. The target variable in the dataset is "Suggested Job Role", making the problem a multi-class classification task. The dataset contains multiple career categories and reflects real-world diversity in student profiles. Proper preprocessing and feature engineering were applied to transform the raw data into a suitable format for machine learning model training and evaluation.

Chapter 4

System Architecture

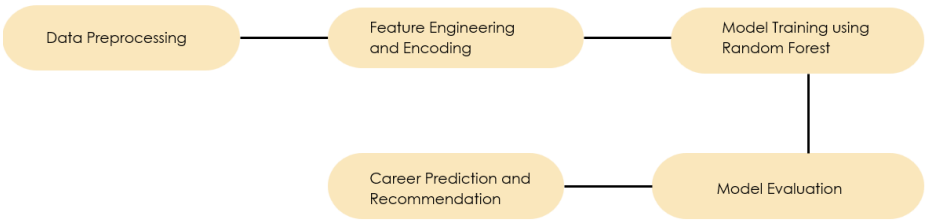


Figure 4.1: System Architecture of Career Prediction System

The architecture consists of the following major components:

- **User Input Module:** Collects student details such as skills, interests, certifications, and academic attributes.
- **Data Preprocessing Module:** Handles missing values, encoding of categorical features, and normalization of numerical data.
- **Feature Engineering Module:** Transforms raw data into meaningful numerical features using techniques such as TF-IDF vectorization.
- **Machine Learning Model:** Applies the trained Random Forest classifier to predict suitable career roles.
- **Prediction Output Module:** Displays the predicted career role along with alternative recommendations through the user interface.

The modular design of the system ensures scalability, maintainability, and consistent performance during both training and deployment.

Chapter 5

Feature Engineering

Feature engineering is a crucial step in this project, as machine learning models require numerical input to perform predictions. The dataset used in this career prediction system contains a combination of numerical, categorical, and textual features. Numerical features such as logical quotient rating, coding skills rating, hackathon participation, and public speaking points were used directly, as the Random Forest algorithm can effectively handle numerical data without scaling.

Categorical features including self-learning capability, teamwork experience, personality traits, reading and writing skills, and memory capability were converted into numerical form using One-Hot Encoding. This method prevents unintended ordinal relationships between categories and improves model interpretability. Textual features such as interested subjects, certifications, workshops, preferred company type, and interested career area were processed using TF-IDF vectorization, which assigns importance to meaningful words while reducing the influence of common terms.

All feature transformations were implemented using a pipeline-based approach to ensure consistent preprocessing during both training and prediction, thereby improving model accuracy and reliability.

Chapter 6

Machine Learning Models and Evaluation

6.1 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm that identifies an optimal hyperplane to separate different career classes. SVM performs well in high-dimensional feature spaces and provides good generalization. In this project, SVM achieved an accuracy of 83%. However, its performance was affected by dataset imbalance and higher computational complexity, making it less suitable for stable deployment compared to ensemble approaches.

6.2 Random Forest Classifier

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and stability. It handles numerical, categorical, and text-based features effectively and is resistant to overfitting. In this project, Random Forest achieved the highest accuracy of 87%, demonstrating strong generalization and robustness across diverse student profiles.

6.3 Model Evaluation and Comparison

The performance of the implemented models was evaluated using accuracy as the primary metric.

Model	Accuracy
Support Vector Machine	83%
Random Forest Classifier	87%

Random Forest was selected as the final model due to its superior accuracy, stability, and ability to handle feature diversity.

Chapter 7

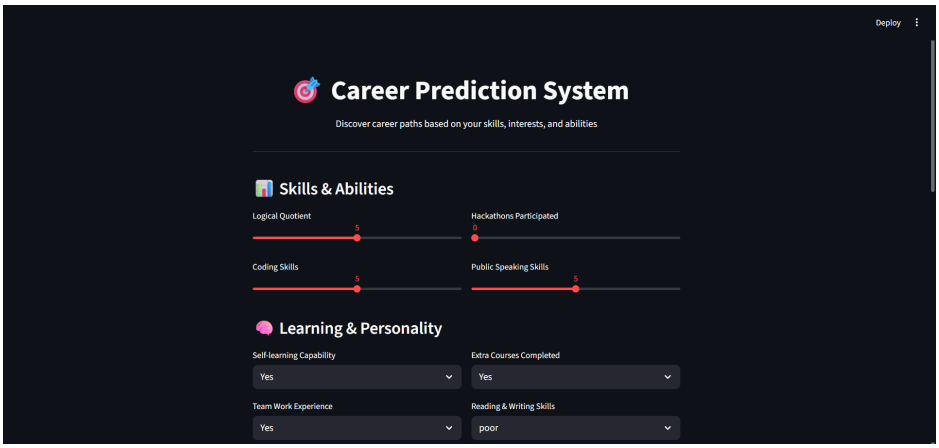
Implementation and User Interface

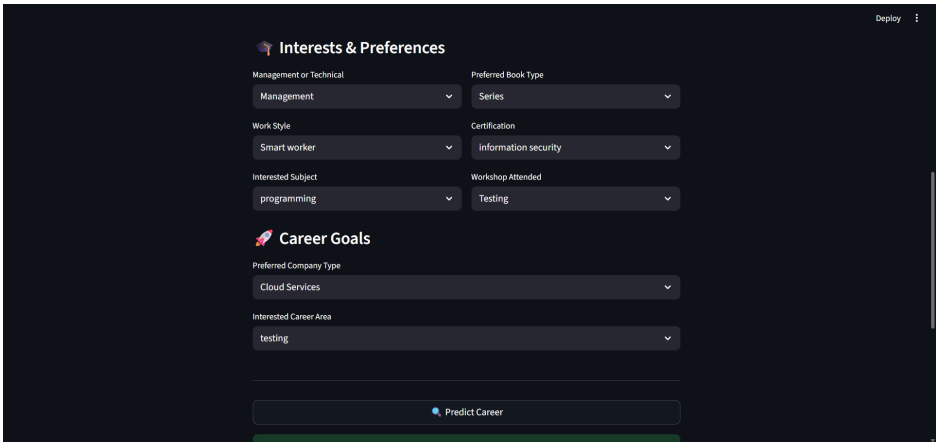
7.1 Implementation Details and Technologies Used:

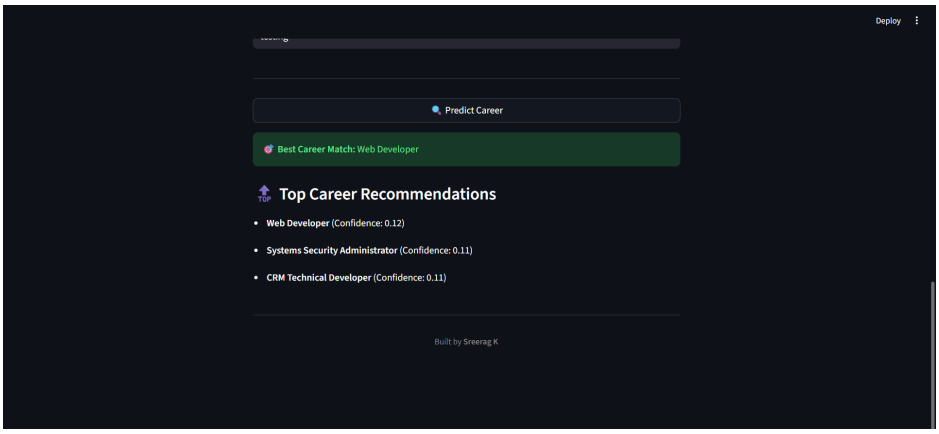
- Programming Language: Python
- Libraries: Pandas, NumPy, Scikit-learn
- Feature Processing: Encoding + TF-IDF
- Model: Random Forest Classifier (Pipeline-based)
- Model Persistence: Pickle

7.2 User Interface Design

A Streamlit-based interface was developed to ensure ease of use and a clean user experience. The UI provides structured input fields, logical grouping of student attributes, and clear output presentation. The output includes a primary career prediction and can be extended to show top recommendations for better decision-making.


Deploy
Career Prediction System
Discover career paths based on your skills, interests, and abilities
Skills & Abilities
Logical Quotient
5
Hackathons Participated
0
Coding Skills
5
Public Speaking Skills
5
Learning & Personality
Self-learning Capability
Yes
Extra Courses Completed
Yes
Team Work Experience
Yes
Reading & Writing Skills
poor


Deploy
Interests & Preferences
Management or Technical
Management
Preferred Book Type
Series
Work Style
Smart worker
Certification
information security
Interested Subject
programming
Workshop Attended
Testing
Career Goals
Preferred Company Type
Cloud Services
Interested Career Area
testing
Predict Career


Deploy
Predict Career
Best Career Match: Web Developer
Top Career Recommendations
Web Developer (Confidence: 0.12)
Systems Security Administrator (Confidence: 0.11)
CRM Technical Developer (Confidence: 0.11)
Built by Sreerag K

Chapter 8

Proposed System

The proposed Career Prediction System is an intelligent, machine learning-based solution designed to assist students in identifying suitable career paths based on their individual profiles. Unlike traditional career guidance approaches that rely on manual counseling and subjective judgment, the proposed system adopts a data-driven methodology to ensure objective and accurate predictions. The system takes various student attributes such as skills, interests, certifications, academic performance, and behavioral characteristics as input and processes them through a structured machine learning pipeline.

A Random Forest classifier is employed as the core prediction model due to its robustness, high accuracy, and ability to handle diverse feature types. The system incorporates effective data preprocessing and feature engineering techniques, including encoding and TF-IDF vectorization, to transform raw input data into meaningful numerical representations. A pipeline-based architecture ensures consistency during both training and prediction phases, minimizing errors and improving reliability.

The proposed system also provides multiple career recommendations instead of a single rigid output, allowing students to explore alternative career options. With a user-friendly Streamlit interface, the system enables easy interaction and real-time predictions, making it a practical, scalable, and efficient tool for modern career guidance.

Chapter 9

Future Scope

The proposed Career Prediction System can be further enhanced in the following ways:

- Integration with real-time job market data to provide career recommendations aligned with current industry demands.
- Salary prediction based on career roles, skills, and experience levels to help students make financially informed decisions.
- Skill gap analysis to identify missing skills and recommend learning paths for achieving desired career goals.
- Implementation of deep learning models to improve prediction accuracy and handle complex feature relationships.
- Deployment as a mobile and web-based application to increase accessibility and reach a wider audience.
- Personalized learning and certification recommendations based on predicted career paths and individual skill profiles.

Chapter 10

Conclusion

This project successfully demonstrates the application of machine learning techniques for career prediction based on student skills, interests, and behavioral attributes. By analyzing a diverse dataset containing numerical, categorical, and textual features, the system provides data-driven career recommendations that support informed decision-making. A pipeline-based architecture was implemented to ensure consistent preprocessing and reliable predictions during both training and deployment. Multiple classification algorithms were evaluated, including Decision Tree and Support Vector Machine, and the Random Forest classifier was selected as the final model due to its superior performance, achieving an accuracy of 87%. The integration of TF-IDF vectorization and One-Hot Encoding enabled effective handling of text and categorical data. The Streamlit-based interface further enhanced usability by allowing users to interactively explore career predictions. Overall, the proposed system offers a scalable and efficient solution for career guidance and can serve as a strong foundation for future enhancements such as real-time job market integration and personalized skill recommendations.