



PROJECT REPORT

Machine Learning  
  
Career Path Prediction System

|  |  |  |  |
| --- | --- | --- | --- |
| **Created By:** | Atharva Vilas Jadhav | **Approved By:** |  |
| **Created On:** | 26-08-2025 | **Approved On:** |  |

Page left blank intentionally

**INDEX**

[**1** **PROJECT DETAILS** 2](#_Toc143445375)

[**2** **SUMMARY** 2](#_Toc143445376)

[**3** **INTRODUCTION** 2](#_Toc143445377)

[3.1 Background 2](#_Toc143445378)

[3.2 Stakeholders 2](#_Toc143445379)

[3.3 Objectives 2](#_Toc143445380)

[**4** **METHODOLOGY** 2](#_Toc143445381)

[4.1 Considerations & Assumption 3](#_Toc143445382)

[4.2 Approach 3](#_Toc143445383)

[4.3 Activities 3](#_Toc143445384)

[**5** **TARGETTED V/S ACHIEVED OUTPUT** 3](#_Toc143445385)

[**6** **CONCLUSION** 3](#_Toc143445386)

[**7** **APPENDICES** 4](#_Toc143445387)

[7.1 Appendix A – Title 4](#_Toc143445388)

**General Instructions for using the Live Project Report Template**

* This template and the subsequent document created using this template is a confidential document and is the intellectual property of Cloud Counselage Pvt. Ltd. Circulating it outside of the organisation without the consent of Cloud Counselage Pvt. Ltd. is the breach of company policies and will lead to legal actions
* This template is a guideline document to communicate the implementation of design ideas and the results of the work to the stakeholders.
* The **text between inequality (< >) is to be replaced** by relevant text
* Please **remove the yellow highlight on the Text** between the inequality (< >). This is done to help you notice the text to be changed/replaced
* The text in *italics* highlighted in grey is just for reference and should be removed after adding the relevant text

# **PROJECT DETAILS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | Career Path Prediction System | | |
| **Project Sponsor** | Tushar Topale | | |
| **Project Manager** | Harshada Topale | | |
| **Start Date** | 23-06-2026 | **Completion Date** | 31-08-2025 |

# **SUMMARY**

The AI-Powered Career Path Prediction System is a web-based tool designed to assist students in making informed career choices. The project addresses the common challenge students face when selecting a career path that aligns with their skills, interests, and academic background. By leveraging a machine learning model, the system analyzes user-provided data—ranging from coding skills to interested subjects—to predict a suitable job role. The final deliverable is an interactive Streamlit application that not only provides an instant career prediction but also suggests relevant online courses to help students begin their professional development journey. This project serves as a proof-of-concept for data-driven student guidance, transforming a traditionally manual counseling process into an automated, scalable solution.

# **INTRODUCTION**

## Background

In today’s fast-paced and highly competitive job market, students face immense pressure to select the right career path at an early stage. This decision is often one of the most crucial turning points in their lives, as it shapes their educational choices, future opportunities, and overall satisfaction in their professional journey. However, many students still make this critical choice based on limited or unreliable sources of information, such as peer influence, societal expectations, or incomplete self-assessment of their own skills and interests. Such decisions, made without proper guidance, can lead to dissatisfaction, underperformance, or frequent career changes later in life, which not only wastes valuable time and resources but also affects long-term confidence.

Traditional career counseling methods, though valuable and insightful, are often dependent on manual processes, face-to-face interactions, and the availability of trained counselors. This makes them difficult to scale and inaccessible to a large segment of students, especially in regions where professional guidance is limited or unavailable. As a result, many young individuals are left without the structured support they need to explore and evaluate career options effectively.

To bridge this gap, there is a growing need for innovative tools that leverage technology to provide personalized, data-driven career insights. A system powered by machine learning can analyze a student’s unique profile—including their academic performance, interests, skills, and personality traits—and recommend career paths that align with both their strengths and aspirations. Such an approach goes beyond generic advice by offering tailored suggestions that maximize the likelihood of long-term success and fulfillment.

This project was initiated with the vision of developing such a system: one that empowers students to make confident, well-informed career decisions. By integrating advanced analytics and machine learning, the project seeks to democratize access to effective career guidance, helping every student—regardless of background—navigate their journey with clarity, purpose, and direction.

## Stakeholders

**1. Students**  
Students are the primary users and direct beneficiaries of the system. By interacting with the platform, they receive **personalized career recommendations** based on their academic performance, skills, interests, and personality traits. This enables them to make more informed and confident decisions regarding their future career paths. For many students, the system acts as a much-needed alternative to traditional counseling, providing them with accessible, data-driven guidance that supports both their academic and professional development.

**2. Cloud Counselage Pvt. Ltd.**  
As the project owner, Cloud Counselage Pvt. Ltd. plays a crucial role in developing, maintaining, and deploying the system. The project not only strengthens the company’s portfolio but also demonstrates its ability to apply **artificial intelligence and machine learning solutions** to real-world challenges in the education and career development sector. Beyond student support, this system also serves as a strategic initiative for **brand visibility, innovation, and client engagement**, positioning Cloud Counselage as a forward-thinking organization committed to empowering the next generation.

**3. Academic Mentors and Counselors**  
Mentors, teachers, and professional career counselors can utilize the system as a **supplementary tool** to enhance their guidance process. While human insight remains irreplaceable, the system provides a strong analytical foundation, offering mentors valuable data and recommendations that can be discussed further with students. This not only improves the effectiveness of counseling sessions but also allows mentors to focus on areas where students need more personal, emotional, or motivational support.

## Objectives

# **Develop a Robust Machine Learning Model** The primary goal was to design and train a reliable machine learning model capable of accurately predicting a student’s potential career path. The model leverages a diverse set of input features—such as academic performance, interests, skills, and personality traits—to ensure predictions are both data-driven and personalized.

# **Design a User-Friendly Web Application** To ensure accessibility and ease of use, the project aimed to build an intuitive web-based platform using **Streamlit**. The application provides students with a seamless interface to enter their details, eliminating technical barriers and allowing for effortless interaction with the system.

# **Integrate the Model for Real-Time Predictions** Another key objective was to embed the trained machine learning model into the Streamlit application, enabling **instant predictions** of career paths. This ensures that students receive immediate feedback and actionable recommendations without delays.

# **Provide Actionable Career Guidance** Beyond predictions, the project emphasizes practical guidance by suggesting **relevant online courses** aligned with the recommended career path. This bridges the gap between insights and implementation, helping students take proactive steps toward skill development and career readiness.

# **Dataset Analysis and Visualization** To enhance understanding and transparency, the project also focused on analyzing and visualizing the underlying dataset. This step uncovers critical insights into the factors influencing career choices and provides a clearer perspective on the relationships between student attributes and career outcomes.

# **METHODOLOGY**

These conventions are all about the positions of line breaks, how many characters should go on a line, and everything in between.

## Considerations & Assumption

1. **Data Quality**  
   The accuracy and reliability of the machine learning model are heavily dependent on the quality and integrity of the dataset used for training—specifically the *PS2\_Dataset.csv*. It is assumed that the dataset is free from significant errors, missing values, or biases that could otherwise negatively impact the model’s performance.
2. **User Honesty**  
   The system assumes that users provide **truthful and accurate information**, particularly for subjective inputs such as self-rated skills, interests, and preferences. Any discrepancies or exaggerations in user responses could affect the precision of the career recommendations.
3. **Static Dataset**  
   Since the model is trained on a **fixed dataset**, it does not dynamically adapt to evolving job market conditions, emerging career roles, or rapidly changing industry demands. It is assumed that the dataset represents a sufficiently relevant and stable snapshot of current career pathways.
4. **Technical Environment**  
   The project assumes access to a **standard Python environment** equipped with essential open-source libraries such as Streamlit, Pandas, Scikit-learn, and others required for machine learning and data visualization. It is also assumed that users will have the computational resources necessary to run the application smoothly.

## Approach

The project was executed following a structured machine learning workflow, from data exploration to final application deployment.

1. **Data Preprocessing & Cleaning:** The initial dataset contained many "dirty" categorical features with free-form text. A crucial first step was to clean and standardize this data. A mapping system was developed to group similar, messy text entries into clean, consistent categories (e.g., "programming" and "coding" were both standardized to "Programming"). This ensured that the data used for training was clean and directly corresponded to the options presented in the final user interface.
2. **Feature Engineering (One-Hot Encoding):** To make the categorical data usable by a machine learning model, **One-Hot Encoding** was employed. This technique converts categorical variables into a numerical format that the model can understand without making incorrect assumptions about the relationships between categories.
3. **Model Selection & Training:** After initial experiments with more complex models like XGBoost resulted in severe overfitting, a **RandomForestClassifier** was chosen. This model is powerful yet inherently more robust and less prone to memorizing the training data, making it a better fit for this specific dataset. The model was trained on a designated training split of the data.
4. **Feature Selection:** A key challenge was the high number of features, many of which were more noise than signal. To address this, a two-stage training process was implemented. A preliminary model was first trained to identify the **top 10 most influential features**. A final, more focused model was then trained using only these top 10 features. This forced the model to learn from the most important signals in the data, drastically reducing overfitting and improving its ability to generalize.
5. **Model Evaluation:** The model's performance was evaluated on a separate, unseen test set. Key metrics from the **classification report**, such as precision, recall, and F1-score for each career path, were analyzed to understand the model's strengths and weaknesses.
6. **Web Application Development:** An interactive web application was built using the **Streamlit** library. The UI was designed to be intuitive, using radio buttons for numerical ratings and dropdown menus for categorical selections, ensuring a smooth user experience. The application includes data visualizations and model insights to provide additional context.

## Activities

The project was broken down into the following key activities:

* **Requirement Gathering:** Analyzed the problem statement, workflow diagrams, and dataset to define the project scope and objectives.
* **Data Exploration (EDA):** Used Plotly to create interactive charts (histograms, box plots, etc.) to understand the distributions and relationships within the data.
* **Data Cleaning:** Implemented functions to standardize and clean the raw categorical data.
* **Model Development:** Built a machine learning pipeline that included data scaling, feature selection, and model training using RandomForestClassifier.
* **Model Evaluation & Iteration:** Repeatedly tested the model, identified overfitting issues, and refined the approach by implementing feature selection to improve performance.
* **UI Development:** Built the front-end of the application using Streamlit, focusing on creating an intuitive and user-friendly interface.
* **Integration:** Integrated the trained model into the Streamlit application to enable real-time predictions.
* **Documentation:** Created all required project documents, including the Project Charter, SRS, and this final report.

# **TARGETTED V/S ACHIEVED OUTPUT**

**1. Targeted Output**  
The core objective of the project was to design and implement a **highly accurate machine learning model** integrated into a user-friendly web application. The envisioned system aimed to reliably predict a student’s ideal career path based on their input features and provide **actionable recommendations**, such as relevant online courses, to guide skill development and career readiness.

**2. Achieved Output**  
The project successfully delivered a **fully functional and interactive web application** built using Streamlit. The application allows students to seamlessly input their details, processes the data through the trained model, and provides career path predictions along with corresponding course suggestions.

A critical milestone in achieving this output was the **implementation of an effective feature selection strategy**, which resolved severe overfitting issues observed in the early stages of development. This resulted in a more stable, generalized model that could perform consistently across varied inputs.

However, the final **predictive accuracy** of the system remains constrained by certain factors—primarily the **limitations and noise within the provided dataset**. While the application demonstrates strong potential and serves as an excellent **proof-of-concept**, its real-world reliability would benefit significantly from training on a **larger, more detailed, and professionally curated dataset**.

# **CONCLUSION**

This project successfully delivered a **complete, end-to-end machine learning application** designed to address one of the most pressing challenges faced by students—making informed career choices. The development journey, starting from raw data preprocessing to deploying a fully interactive web application, involved navigating several real-world challenges. Among these, **data quality issues and model overfitting** were the most significant hurdles, which were effectively mitigated through robust feature selection techniques and iterative experimentation.

A key takeaway from the project is that **the quality, relevance, and richness of input data are the most decisive factors** in determining the overall success of any machine learning model. Despite working within the limitations of the provided dataset, the final system—powered by a **RandomForestClassifier integrated with a carefully designed feature selection pipeline**—proved to be a strong **proof-of-concept**, demonstrating the feasibility of data-driven career guidance.

For future development, the most impactful recommendation would be to focus on **data enrichment and diversification**. Expanding the dataset to include more granular details such as specific project experiences, subject-wise academic performance, extracurricular achievements, and psychometric or aptitude assessments would significantly enhance the predictive accuracy and personalization of the model.

In its current state, the system already serves as a **valuable, accessible, and data-driven starting point** for students embarking on their career exploration journey. With continued refinement and richer datasets, it has the potential to evolve into a **highly reliable and scalable career guidance platform** that can support students across diverse educational and socio-economic backgrounds.

# **APPENDICES**

## Appendix A – Title

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Deliverable Name** | **Status** |
| 1 | Development Log | Completed |
| 2 | GitHub Repository/Source Code | Completed |
| 3 | Lessons Learned Log | Completed |
| 4 | Project Report | Completed |
| 5 | Project Schedule | Completed |
| 6 | RAID Log | Completed |
| 7 | Project Video | Completed |
| 8 | Work Breakdown Structure | Completed |