CareerPath AI - Final Product Report

CareerPath AI Team

December 2, 2025

Members

- Julián Andrés León Pabón
- Ryan Mintoo
- Andrew Nguyen

1 VGC Group Project Proposal

Tips: We strongly recommend you start early to discuss your proposal with your teammate to make sure you are on track to meet the assignments' requirements early on.

Title

CareerPath AI: Educational Data-Driven Career Trajectory Prediction System

1.1 Honor Code and LLM Usage for this Proposal (5pts)

Include a statement affirming adherence to the honor code and ethical use of Large Language Models (LLMs) for this project.

Honor Code and LLM Usage Statement

We, the members of Group 15 (Ryan, Andrew, and Julian), affirm our full commitment to upholding the academic honor code throughout this project. We pledge to produce original work that reflects our genuine understanding, effort, and learning.

Regarding the use of Large Language Models (LLMs), we will employ these tools ethically and responsibly as assistive resources rather than substitutes for our own critical thinking and work. Specifically, we will use LLMs (such as ChatGPT, Claude, and GitHub Copilot) for:

- **Research assistance**: Clarifying ML concepts and exploring different algorithmic approaches.
- **Code scaffolding**: Generating boilerplate code and suggesting implementation patterns.
- **Debugging support**: Identifying errors and understanding error messages.
- **Documentation**: Improving code comments and technical writing clarity.

We commit to:

- **Verify all outputs**: Critically evaluating and testing any LLM-generated code or content.

- **Understand before using**: Ensuring we comprehend any code or concepts before incorporating them.
- **Proper attribution**: Clearly documenting when LLMs significantly contributed to our work.
- **Original synthesis**: Using our own judgment to integrate, modify, and adapt AI suggestions to fit our project goals.

This project represents our own learning journey, with LLMs serving as educational tools to enhance our understanding and development of machine learning skills.

1.2 Learning Objectives (10pts)

List at least two learning objectives for your project. These should include both conceptual understanding and skill development.

1. **Learning Objective 1 (Conceptual Understanding)**: Understand how machine learning models can be ethically applied to predict career outcomes while recognizing their limitations and biases in capturing non-quantifiable success factors such as personality traits, cultural fit, and personal preferences.
2. **Learning Objective 2 (Skill Development)**: Develop proficiency in data preprocessing, feature engineering, and supervised learning techniques (regression and classification algorithms) to build a predictive model for career trajectory analysis using Python and relevant ML libraries.

1.3 Timeline (10pts)

Outline how you plan to spend 5-10 hours on your project. Break down the time into specific tasks or milestones.

- **Hour 1-2: Research and gather resources**
  Research existing career prediction tools and publicly available career datasets in CSV format (from Kaggle, government labor statistics). Review tutorial articles on basic machine learning for career prediction. Identify key features from educational data (GPA, major, university type, extracurricular activities) that could predict career paths. Define 3-5 specific career categories to predict and establish success metrics.
- **Hour 3-4: Design the project structure and plan**
  Design the overall system architecture including data pipeline, model selection, and output format. Create a project plan outlining data sources, preprocessing steps, and ML algorithms to test (logistic regression, random forests, neural networks). Develop user stories and define the interface for how users will input their educational data and receive career predictions.
- **Hour 5-6: Start coding the basic functionalities**
  Load and preprocess data from CSV files using pandas. Clean the dataset by handling missing values and normalizing features. Perform basic exploratory data analysis with visualizations (matplotlib/seaborn) to understand patterns. Create additional features from existing data (e.g., categorize majors into STEM/Business/Arts). Begin implementing the prediction model using scikit-learn with a simple train-test split approach.

- **Hour 7-8: Test and debug the initial version**
  Train multiple ML models and evaluate performance using appropriate metrics (accuracy, precision, recall, F1-score). Implement cross-validation to ensure model generalizability. Test the model with sample educational profiles and validate predictions against known career outcomes. Debug any data processing issues and refine feature selection based on model performance and feature importance analysis.

- **Hour 9-10: Refine and add advanced features**
  Address ethical considerations by implementing bias detection mechanisms and providing transparency about model limitations. Develop a user-friendly interface for inputting educational data and displaying predictions with confidence intervals. Create documentation explaining the model's capabilities, limitations, and ethical considerations. Prepare final presentation materials and conduct end-to-end testing of the complete system.

2 Final Product Description (55pts)

2.1 Proposed Versions

- **Minimum Viable Product (MVP)**: A basic Streamlit application using a synthetic dataset and a simple Random Forest classifier to predict career paths based on user inputs (OCEAN traits and aptitude scores).
- **Target Product**: The current version. It integrates a hybrid dataset (Real Kaggle data + Synthetic augmentation) for better accuracy (71.6%). It features a full Model Context Protocol (MCP) server, OpenAI GPT-5 integration for personalized advice, and a polished UI with interactive visualizations.
- **Reach Version**: A fully deployed cloud application (AWS/Heroku) with user authentication, history tracking, a mobile-responsive design, and advanced ensemble models (XGBoost/Neural Networks) for higher precision.

2.2 Description of Final Product

CareerPath AI is an intelligent career guidance system designed to help students make data-driven career decisions. The target audience includes high school and college students seeking personalized career advice. The problem it addresses is the lack of objective, personalized guidance in traditional career counseling. Key features include real-time career predictions based on personality (OCEAN) and aptitude, interactive radar charts, and AI-powered explanations generated by GPT-5. Technically, the project uses Python, Scikit-learn for machine learning, Streamlit for the web interface, and implements a Model Context Protocol (MCP) server for modularity. It leverages a hybrid dataset of real and synthetic records to ensure robust and fair predictions.

2.3 Video Demo

Link to Video Demo: `https://youtu.be/CU1qiY_81KI`

2.4 Input and Coding Files

The following key files are included in the submission:

- `web/app.py`: Main Streamlit web application.
- `src/models/train.py`: Script for training the Machine Learning models.
- `src/data/preprocess.py`: Data cleaning and preprocessing pipeline.
- `src/experiments/bias_demo.py`: Script demonstrating ethical bias testing.
- `src/mcp/career_data_server.py`: Implementation of the MCP server.
- `README.md`: Instructions for running the project.
- `PROJECT_SUMMARY.md`: Detailed project documentation.

3 Consultation and Use of LLMs (10pts)

3.1 Consultation Description

I consulted with my teacher, Daniel, specifically regarding the ethical considerations of the project. His feedback was crucial in shaping the "Ethical Considerations" section, particularly in defining the scope of bias testing and the importance of addressing cross-cultural implications between the U.S. and Ecuador.

3.2 Use of LLMs

Large Language Models (specifically Gemini and GPT-5) were utilized extensively throughout the development process:

- **Coding**: LLMs assisted in generating boilerplate code for the Streamlit interface and the MCP server structure.
- **Debugging**: They were used to troubleshoot errors in the data pipeline, specifically with Pandas DataFrame operations and Scikit-learn version compatibility.
- **Idea Generation**: LLMs helped brainstorm the "Bias Demonstration" experiment to visually show the impact of skewed training data.

4 Ethical Considerations (10pts)

4.1 Data Provenance and Consent

The project uses a hybrid dataset. The core data comes from the **Kaggle Career Prediction Dataset** (Open License), which contains anonymized public data. We augmented this with synthetic data generated based on statistical distributions. No personal data from real users is stored; input data is processed in-memory and discarded after the session.

4.2 Privacy and Security Risks

Since the application runs locally, privacy risks are minimal. However, if deployed, sensitive attributes like "Neuroticism" scores could be exposed. We mitigate this by not implementing permanent database storage for user profiles in this version.

4.3 Fairness and Bias

We identified that the model could be biased against individuals with lower aptitude scores if the training data is skewed. To mitigate this, we implemented a **Bias Demonstration** (`src/experiments/bias_demo.py`) to audit the model. We also used stratified sampling to ensure all career categories are equally represented in the training set.

4.4 Misuse and Safety

Potential misuse includes using the tool as a definitive "gatekeeper" for career choices. To prevent this, we include clear disclaimers stating that predictions are for guidance only and should not replace professional counseling.

4.5 Transparency and Accountability

We provide full transparency through open-source code and a detailed `README.md`. The model's limitations (e.g., 71.6% accuracy) are clearly communicated to the user in the "About" section of the app.

4.6 Cross-Cultural and Accessibility (US vs. Ecuador)

- **Cross-Cultural**: The current model is trained on data that may reflect Western/US career norms. In Ecuador, career paths might be more influenced by family tradition or local economic factors which the OCEAN model might not fully capture. Future versions should include locally sourced data.
- **Accessibility**: The app is currently in English. For the Ecuadorian context, a Spanish localization is essential. We designed the UI with clear icons and simple language to be accessible to non-native English speakers.

4.7 Intellectual Property

The code is licensed under the MIT License, allowing for open use and modification. All external libraries (Scikit-learn, Streamlit) are used in compliance with their respective licenses.