Spaceship Titanic Prediction Project

Overview

This project focuses on predicting whether passengers aboard the fictional Spaceship Titanic were "transported" to an alternate dimension, based on a dataset from a Kaggle binary classification competition. The goal was to build a machine learning model capable of accurately classifying passenger outcomes using various features like demographics, travel arrangements, and spending habits.

Project Structure

- `main.py`: The main Python script containing the entire data loading, preprocessing, model training, evaluation, and submission generation pipeline.
- `README.md`: This file, providing an overview and instructions.
- `train.csv`: The training dataset (obtained via Kaggle API).
- `test.csv`: The test dataset (obtained via Kaggle API).
- `submission.csv`: The generated prediction file for Kaggle submission.

Setup and Installation

To run this project, you'll need Python and a few libraries.

1. **Python:** Ensure you have Python 3.x installed.
2. **Install Libraries:** It's recommended to use a virtual environment.
   ```
   # (Optional) Create and activate a virtual environment
   python -m venv venv
   # On Windows:
   .\venv\Scripts\activate
   # On macOS/Linux:
   source venv/bin/activate

   # Install required libraries
   pip install pandas numpy scikit-learn kaggle
   ```
3. **Kaggle API Credentials (Optional for Running, Required for API Download):** If you wish to download the data programmatically using the Kaggle API (as done initially in this project's development), ensure you have your `kaggle.json` file placed in `~/.kaggle/` (on Linux/macOS) or `C:\Users\<Windows-username>\.kaggle\` (on Windows). You can generate this file from your Kaggle account settings.

Data Acquisition

The `train.csv` and `test.csv` datasets are programmatically obtained using the Kaggle API. The `main.py` script is designed to access these files from local storage paths once they have been downloaded. Ensure these files are present in the project directory or the specified local paths.

How to Run the Project

1. **Place Data Files:** Ensure `train.csv` and `test.csv` (obtained via Kaggle API) are in the same directory as `main.py`, or update the `pd.read_csv` paths in `main.py` to point to their correct locations.
2. **Execute the Script:** Open your terminal or command prompt, navigate to the project directory, and run:
   `python main.py`
3. **Output:** The script will print validation accuracy to the console and generate a `submission.csv` file in the same directory.

Model Performance

- **Validation Accuracy:** Approximately 0.797
- **Kaggle Submission Score:** 0.79424

These scores indicate a strong performance, successfully exceeding the project's target threshold.

Dependencies

- `pandas`
- `numpy`
- `scikit-learn`
- `kaggle` (for API functionality, if used for data download)

Author

Ellie Capra