Machine Learning Project Documentation

# Project Title: Predicting Olympic Medal Counts

## Overview

This project aims to build a machine learning model to predict the number of medals a country will win in the Olympic Games based on historical and current data. The dataset used for this project is sourced from historical Olympic games data.

## Project Steps

1. Form a Hypothesis

* **Objective**: To predict the number of medals a country will win in the Olympic Games.
* **Hypothesis**: We hypothesize that various factors, including a country's historical performance, population, and economic indicators, influence its medal count in the Olympics.

2. Data Collection and Exploration

* **Data Source**: The dataset used in this project is obtained from [Kaggle](https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results).
* **Data Files**:
  + **teams.csv**: Team-level data used for analysis and modeling.
  + **athlete\_events.csv**: Original athlete-level data.
* **Exploration**: Explore the dataset to understand its structure, variables, and relationships between different features.

3. Data Reshaping (if necessary)

* **Reshaping Data**: If required, reshape the data to predict the target variable (medal count). This may involve aggregating athlete-level data into team-level data.

4. Data Cleaning

* **Data Cleaning**: Prepare the data for machine learning by handling missing values, outliers, and ensuring data consistency.

5. Error Metric

* **Error Metric**: Select an appropriate error metric to measure the performance of the machine learning model. Common metrics include Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE).

6. Data Splitting

* **Data Splitting**: Split the dataset into training and testing sets to evaluate the model's performance. Common splits include 80% for training and 20% for testing.

7. Model Training

* **Model Selection**: Choose a suitable machine learning algorithm for regression tasks. Options include linear regression, decision trees, random forests, or gradient boosting.
* **Feature Engineering**: Create relevant features based on historical and current data.
* **Model Training**: Train the chosen model on the training data.
* **Model Evaluation**: Evaluate the model's performance on the testing data using the selected error metric.

## Code

The project code is organized into two Jupyter Notebook files:

1. **machine\_learning.ipynb**: This notebook contains the main code for the machine learning project, including data preprocessing, model training, and evaluation.
2. **data\_prep.ipynb**: This notebook contains the code used to generate the team-level dataset from the original athlete-level dataset.

## Local Setup

Installation

To run this project locally, you need the following:

* Python 3.8+
* Required Python packages:
  + pandas
  + numpy
  + scikit-learn
  + seaborn

## Data

You can download the required data files from the following links:

* [teams.csv](https://drive.google.com/uc?export=download&id=1L3YAlts8tijccIndVPB-mOsRpEpVawk7): Team-level data for the project.
* [athlete\_events.csv](https://drive.google.com/uc?export=download&id=1Ah4wOyNFMGREq8Yw_Jbv7u2CeI_6tpn5): The original athlete-level data.

Ensure that you have the data files in the same directory as the project notebooks for the code to work correctly.

This project documentation provides an overview of the objectives, steps, and resources needed to replicate and understand the machine learning project for predicting Olympic medal counts.

Top of Form