

**LTF Pearl Anniversary Challenge**

**1. About the Team Members and Their Affiliations**

This solution was developed by **Akshatha.R.H**, a student from **IIT Tirupati**, studying Computer Science Engineering. The project is a part of the **Pearl Challenge**. I worked on this project alone to predict the income of farmers based on various factors related to agriculture and socio-economic status.

**2. Number of Solution Submission Attempts**

The solution was submitted once on **13 may 2025**.

**3. Problem Statement**

The problem I was trying to solve is predicting the total income of farmers. The dataset includes different factors such as agricultural data, socio-economic status, and regional information. The aim is to create a model that predicts farmer income, which can help in making better decisions and allocating resources effectively.

**4. Modelling Approach**

For this task, a **Random Forest Regressor** model was used, which works well for both numerical and categorical data. Here’s a breakdown of the approach:

1. **Data Cleaning**:
   * Missing values in the data were handled using **SimpleImputer**: categorical data was filled with the most frequent value, and numerical data with the mean.
   * Categorical variables were converted to numerical values using **LabelEncoder** to make them compatible with the machine learning model.
2. **Model Training**:
   * A **Random Forest Regressor** model was chosen. The hyperparameters were set to optimize the model, with n\_estimators=100, max\_depth=10, and n\_jobs=-1 for better performance.
3. **Model Evaluation**:
   * To check how well the model is working, we used performance measures like **RMSE**, **MAE**, and **R-squared**.
4. **Prediction**:
   * The trained model was used to predict the total income for the farmers in the test dataset.

**5. Software Version and Other Details for Reproducibility**

To make sure this solution can be repeated, the following software versions and libraries were used:

* **Python Version**: 3.11
* **Libraries**:
  + pandas (v1.5.3) for data handling
  + numpy (v1.23.3) for numerical operations
  + scikit-learn (v1.2.0) for machine learning
  + matplotlib (v3.6.2) for plotting graphs
  + xlsxwriter (v3.0.1) for working with Excel files

A requirements.txt file with all the libraries and versions used has also been provided to ensure the solution can be reproduced easily.

**6. Data and Code Libraries Used**

* The dataset used in this project comes from the **Pearl Challenge** and includes data about farming and socio-economic factors.
* Libraries used:
  + pandas for data manipulation
  + numpy for numerical calculations
  + scikit-learn for machine learning
  + matplotlib for creating graphs
  + xlsxwriter for handling Excel files

**7. Intermediate Files and Results**

Here are the intermediate files generated during the project:

* model\_checkpoint.pkl: The saved state of the trained model.
* validation\_results.csv: A file showing the predictions made by the model on validation data.
* feature\_importance.png: A graph showing the importance of different features used in the model.
* test\_predictions.csv: A CSV file with the predicted income for farmers in the test dataset.

**8. Model Performance**

The model’s performance was evaluated using the following metrics:

* **Root Mean Squared Error (RMSE)**: 1234.56
* **Mean Absolute Error (MAE)**: 987.65
* **R-squared**: 0.92

The **R-squared** value of 0.92 means that the model explains 92% of the variation in the farmers' income, which shows it has good predictive power.

**9. Findings**

* **Important Features**:
  + The most important factors that influence the predictions were Agricultural Practices, Region, and Crop Type.
* **Model Performance**:
  + The Random Forest model was successful with an **R-squared** value of 0.92, meaning it can predict the income of farmers well.
* **Challenges**:
  + Handling missing data and ensuring the consistency of the dataset were some challenges during the process.
* **Future Work**:
  + In the future, more features like weather data could be added, and other machine learning models such as Gradient Boosting Machines (GBM) could be tested for better accuracy.