Task - Tech Challenge
Amay Chivate sandeep.amay21@vit.edu

Objective:
Build a basic machine learning model to predict house prices using the Boston Housing Dataset (or any other public dataset like Titanic Survival or Iris Classification, depending on preference).
Task Details:
1. Load the dataset using Python (preferred: pandas, scikit-learn).
2. Perform basic data cleaning and preprocessing.
3. Do exploratory data analysis (EDA) – visualize key relationships.
4. Split the data into training and test sets.
5. Train a simple regression/classification model (Linear Regression, Decision Tree, or Random Forest).
6. Evaluate the model using suitable metrics (e.g., RMSE, Accuracy, etc.).
7. Share the code on GitHub and send the link with a brief explanation (max 300 words) on:
○ What you have done
○ Why you have chosen a specific model
○ Any challenges you have faced
Submission Time: 2–3 days

Github Link - https://github.com/rttr-7/Task-California-

(Note: Since the Boston housing dataset was removed from scikit-learn due to ethical concerns, I used the California housing dataset as an alternative for this analysis.)

I developed a Random Forest regression model to predict median house values using the California Housing dataset. After performing data cleaning and exploratory data analysis (EDA), I split the data into training and testing sets. The model was trained on various features such as median income, house age, average rooms, and geographic coordinates. I evaluated its performance using metrics like RMSE, MAE, and R^2 score, ensuring reliable predictions. To demonstrate a real-world case, I tested the model with a custom input case and successfully generated predictions.
Due to the dataset's inherent characteristics, particularly non-linear relationships between features and home values, presence of skewed distributions, and complex spatial patterns, a Random Forest regression model was selected. This approach effectively handles such complexities without requiring extensive data transformation. The model demonstrates strong predictive capability while naturally managing the dataset's challenges.

One of the challenges I faced was managing the large size of the trained Random Forest model file (~138 MB).

Key technical aspects include:
Data visualization of distributions and relationships
Feature engineering addressing skewed variables
Model training with built-in handling of non-linear patterns