**Name: Rishav Chandra Acharya**
**Email: rishav.c.acharya@vanderbilt.edu**
**VUNetID: acharyrc**

## Background Information

The housing market in California has always been a key area of interest due to the dynamic fluctuations in housing prices across the state. Accurate predictions of housing prices are essential for buyers, sellers, investors, and policymakers to make informed decisions. To address this, machine learning models can be employed to predict housing prices based on various features such as median income, housing median age, total rooms, and population. This project utilizes the California Housing dataset to build predictive models and evaluate their performance in predicting the median house value.

**Dataset:** The dataset chosen is related to California housing prices. It contains columns like longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, and median house value.

https://raw.githubusercontent.com/dataprofessor/data/refs/heads/master/california_housing_test.csv

## Problem Statement

The objective of this project is to predict the 'Median House Value' for districts in California based on the features provided in the dataset. The problem falls under the category of regression, as the goal is to predict a continuous variable.

## Hypothesis

It is hypothesized that Random Forest Regressor, being a more complex and flexible model, will outperform Linear Regression in predicting house prices, as it can capture non-linear relationships and interactions between features that a linear model might miss.

## Methods

Two algorithms were chosen to tackle this regression problem: Linear Regression and Random Forest Regressor.

- Linear Regression: This model assumes a linear relationship between the independent variables and the target variable. It fits a straight line through the data to minimize the difference between the predicted and actual values (based on the

least squares criterion). Linear regression is simple, fast, and interpretable but can struggle when the relationship between variables is non-linear.

- Random Forest Regressor: This is an ensemble learning method that builds multiple decision trees and averages their predictions. Random Forest is more flexible, capable of capturing complex, non-linear relationships and interactions between variables, and less prone to overfitting compared to a single decision tree. However, it can be slower to train and less interpretable than linear models.

## Data Cleaning and Preprocessing:

The data was cleaned by:

- Deleting the duplicated rows
- Filling the empty values with the mean value of the columns
- Normalizing Median Income using Z-score to adjust its scale (the scale is unknown).
- Removing Outliers (case: if Median Income is more than 3 standard deviations from the mean).

**Train-Test Split**: The data was split into a training set (80%) and a test set (20%) to evaluate model performance.

**Evaluation Metric**: The performance of both models was evaluated using the Mean Squared Error (MSE) and $R^2$ (coefficient of determination) on both the training and test sets. The $R^2$ score provides insight into the proportion of variance in the target variable that can be explained by the independent variables.

**Results and Discussion**

**Linear Regression:**

- Training Mean Squared Error: 4,787,086,799.53
- Training $R^2$: 0.6282
- Test Mean Squared Error: 5,157,558,978.37
- Test $R^2$: 0.5854

Linear Regression provided a decent performance, explaining about 58.54% of the variance in the test data. However, the model seems to have a moderate error rate in predicting house prices, as shown by the high Mean Squared Error (MSE). The performance on the test set is slightly worse than on the training set, indicating slight overfitting, but overall, the model generalizes reasonably well.

**Random Forest Regressor**:

- Training Mean Squared Error: 4,787,086,799.53
- Training $R^2$: 0.6282
- Test Mean Squared Error: 6,782,409,436.95
- Test $R^2$: 0.4548

The Random Forest model performed worse than expected, especially on the test set, where it achieved a lower $R^2$ score of 45.48%. This suggests that while the model can capture more complex relationships in the data, it might have overfit the training data, leading to a poorer performance when tested on unseen data. The model's higher MSE also indicates that its predictions were less accurate compared to Linear Regression.

**Discussion:**

The results demonstrate that Linear Regression outperformed Random Forest Regressor in terms of test set performance. This was somewhat surprising, as Random Forest Regressor is typically expected to perform better on complex datasets due to its ability to capture non-linear relationships. However, the performance discrepancy could be attributed to the relatively small size of the dataset or the tuning of the Random Forest model. Additionally, since the training $R^2$ values for both models were identical, this suggests that further tuning (such as adjusting hyperparameters for Random Forest) could potentially improve the model's performance.

**Conclusion**

In this project, both Linear Regression and Random Forest were applied to predict California housing prices. Contrary to expectations, Linear Regression provided better results on the test data, with a higher $R^2$ and lower MSE. Further exploration and tuning of the Random Forest model, such as adjusting the number of trees or the maximum depth, could lead to improved performance.

**Outside Resources Used**

**Scikit-learn Documentation**: For implementing both Linear Regression and Random Forest Regressor algorithms, as well as splitting the dataset and evaluating model performance.

**Pandas Library:** It was used for data manipulation and cleaning.

**NumPy Library:** It was used for numerical computations, particularly in calculating metrics such as MSE.