Boston Housing Dataset Analysis

(a) Definition of the Problem:

The Boston Housing Dataset is a collection of data concerning various factors affecting housing prices in Boston. The dataset includes 506 instances, each with 13 features such as crime rate, average number of rooms, and accessibility to highways, among others. The goal is to predict the median value of owner-occupied homes in thousands of dollars.

(b) Rationale of Target Variable Selection:

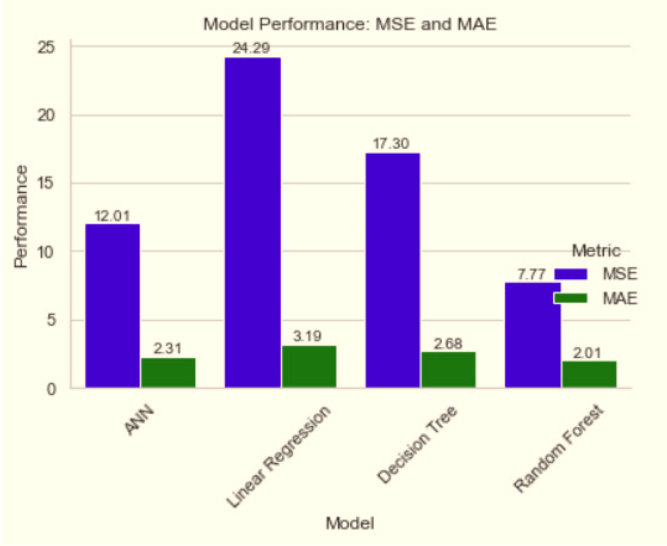
The **median value of owner-occupied homes** is a crucial indicator of the housing market's health and performance. It represents the central point around which housing prices revolve, giving potential buyers and sellers a clear idea of the price range they can expect.

(c) Suitable Machine Learning Approach:

This problem is a **regression task, as the goal is to predict a continuous numerical value** (median housing price) based on a set of features. Algorithms like linear regression, decision trees, and random forests are appropriate for this problem. Due to the relatively small dataset size and the need to handle features with varying scales, ensemble methods like random forests are particularly effective.

(d) Conclusion with Visual:

1. In conclusion, the correlation matrix of the Boston Housing Dataset suggests that factors such as status of the population, average number of rooms, pupil-teacher ratio and property tax significantly influence the housing prices.
2. Based on Hyperparameter tuning on Random Forest the important features are status of the population, average number of rooms,distance to Boston employment centers, crime rate and pupil-teacher ratio.
3. Below is the performance comparison plot between
    a. Neural Network
    b. Linear Regression
    c. Decision Tree
    d. Random Forest



(e) Scope for Future Work:

1. **Feature Engineering** can play a crucial role in enhancing model accuracy by adding new features that could capture valuable insights. For example, features that capture the distance to important city landmarks, proximity to public transportation.
2. Experiment with **different ways of splitting your data** into training, validation, and test sets. Cross-validation, time-based splits, or stratified sampling could provide different perspectives on your model's performance.
3. **Bayesian Regression** techniques can provide not only point predictions but also uncertainty estimates for predictions. This can be particularly valuable in real estate predictions, where understanding the uncertainty of predictions is crucial.

(f) Exceeding Expectations and Extra Work:

1. We experimented with advanced regression techniques - **neural networks** to potentially improve the predictive performance. We found that neural networks outperformed linear regression and decision trees, but it was not able to outperform random forests. Few possible explanations for this finding are:
   a. Neural networks are more complex models than decision trees and random forests. Thus, they require more data to train, and are more prone to overfitting.
   b. Neural networks are not as interpretable as decision trees and random forests. This can make it difficult to understand how the model is making its predictions.
2. **Hyperparameter tuning on Random Forest**: Performed an extensive hyperparameter tuning process for a Random Forest model on the Boston Housing Dataset. Implemented GridSearchCV to optimize parameters. This is in-depth analysis aimed to enhance model accuracy and generalization, demonstrating a comprehensive approach.