

Machine Learning Report

I used the California Housing dataset from the skleam.datasets module. The data includes 8 numerical features and 20,640 samples. The source is from the official Scikit- learn repository.

Link : <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html>

This dataset was chosen because it is realistic, easy to understand, and commonly used in ML for both regression and classification.

I framed the target (median house value) in two ways:

* Regression: Predicting the actual value.
* Classification: Grouping values into Low, Medium, and High.

The dataset had no missing values. Preprocessing steps included scaling features using StandardScaler and encoding labels using LabelEncoder.

Data was split into 80% training and 20% testing.

I applied 7 classification models (ANN, SVM, Naive Bayes, KNN, Random Forest, Decision Tree) and 6 regression models including Linear Regression and Random Forest Regressor.

Random Forest Classifier gave the best accuracy (81.44%) in classification, and Random Forest Regressor gave the lowest RMSE (0.5051) in regression.

All results were visualized using confusion matrices, bar charts, and heatmaps. These visualizations are saved in the "Results" folder on GitHub.

Code, data (original and preprocessed), and model outputs are organized in separate

folders for clarity and easy access.

GitHub Repository Link:

<https://github.com/Tarfah-Naif/ML_Project_California>