

BCI Campus

Bachelor of Information technology (Honours)

BSIT 31022- Introduction to Artificial Intelligence

Title

ANALYSIS OF CALIFORNIA HOUSING DATASET

Student Name: M.K.T.Chathuranga

Student ID: BSIT230147

Instructor: Mr.Yohan

Submission Date: 20th October 20

Table of Contents

Contents

Table of Figures.....	2
1. Introduction	3
2. Dataset Description.....	4
3. Methodology.....	4
3.1 Data Preprocessing	4
3.2 Descriptive Statistics	5
3.3 Data Visualization	5
3.4 Model Building	5
3.5 Model Evaluation	5
4. Results and Discussion	6
4.1 Descriptive Statistics Summary.....	6
4.2 Key Visual Insights.....	6
4.3 Model performance	8
4.4 Discussion.....	9
4. Conclusion.....	9
5. References	10

Table of Figures

Figure 1: Histogram of Median House Value	6
Figure 2: Median Income vs Median House Value	7
Figure 3: Geographical Distribution	7
Figure 4: Correlation Matrix.....	8

1. Introduction

The California Housing Dataset (U.S. Census, 1990) contains 20,640 samples and 9 numerical attributes related to housing and demographic characteristics of California districts. This report explores the dataset, visualizes relationships, and builds regression models to predict median house value. Objectives are: exploratory analysis, visualization, and predictive modeling using Linear Regression and Decision Tree Regressor. The California Housing Dataset was collected by the U.S. Census Bureau in 1990 and is widely used for regression and predictive modeling tasks. It contains demographic, economic, and housing-related information across various districts in California.

This dataset includes 20,640 samples and 9 numerical attributes, making it ideal for understanding how socioeconomic and geographical factors influence house prices.

Objective

The main goal of this analysis is to:

- Explore and visualize the dataset.
- Understand relationships between features and median house value.
- Build predictive models (Linear Regression and Decision Tree Regressor).
- Evaluate model performance using statistical metrics.

2. Dataset Description

Source: The dataset is obtained from Scikit-learn's California Housing Dataset.

Attributes Overview:

Table 1: Attributes Overview

Attribute	Description
MedInc	Median income in block group (10k USD units)
HouseAge	Median house age in block group
AveRooms	Average number of rooms per household
AveBedrms	Average number of bedrooms per household
Population	Population of block group
AveOccup	Average number of household members
Latitude	Latitude coordinate
Longitude	Longitude coordinate
MedHouseVal	Median house value (target variable)

- Features vs Target Variable:

Features (Independent Variables): MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude

Target Variable (Dependent): MedHouseVal (Median House Value)

3. Methodology

This section explains the step-by-step analysis and modeling approach.

3.1 Data Preprocessing

- Loaded dataset using Pandas.
- Checked for missing values and removed null entries using (df.dropna).
- Verified data structure and types using (df.info) and previewed the first 10 rows.

3.2 Descriptive Statistics

For selected attributes (median_income, total_rooms, population), computed:

- Mean, Median, Standard Deviation, Minimum, and Maximum values.

These descriptive statistics helped identify the range and distribution of housing-related features.

3.3 Data Visualization

To better understand data patterns, several plots were generated:

- Histogram: Distribution of Median House Value.
- Scatter Plot: Relationship between Median Income and Median House Value.
- Geographical Plot: Longitude vs Latitude showing regional house price variations.
- Heatmap: Correlation matrix to reveal relationships among features.

3.4 Model Building

Two regression models were trained to predict MedHouseVal:

1. Linear Regression:
 - A simple linear model capturing linear relationships between features and house value.
 - Trained using 80% of the data, tested on 20%.
2. Decision Tree Regressor:
 - A non-linear model capturing complex feature interactions.
 - Risk of overfitting mitigated by setting a random state for reproducibility.

3.5 Model Evaluation

Both models were evaluated using:

- Mean Squared Error (MSE): Measures average squared difference between predicted and actual values.
- R² Score: Measures how well model explains variance in the target variable.

4. Results and Discussion

4.1 Descriptive Statistics Summary

Table 2: Descriptive Statistic Summary

Feature	Mean	Median	Std Dev	Min	Max
Median Income	3.87	3.55	1.90	0.49	15.00
Total Rooms	2635.76	2127.00	2181.62	2.00	39320.00
Population	1425.48	1166.00	1132.46	3.00	35682.00

4.2 Key Visual Insights

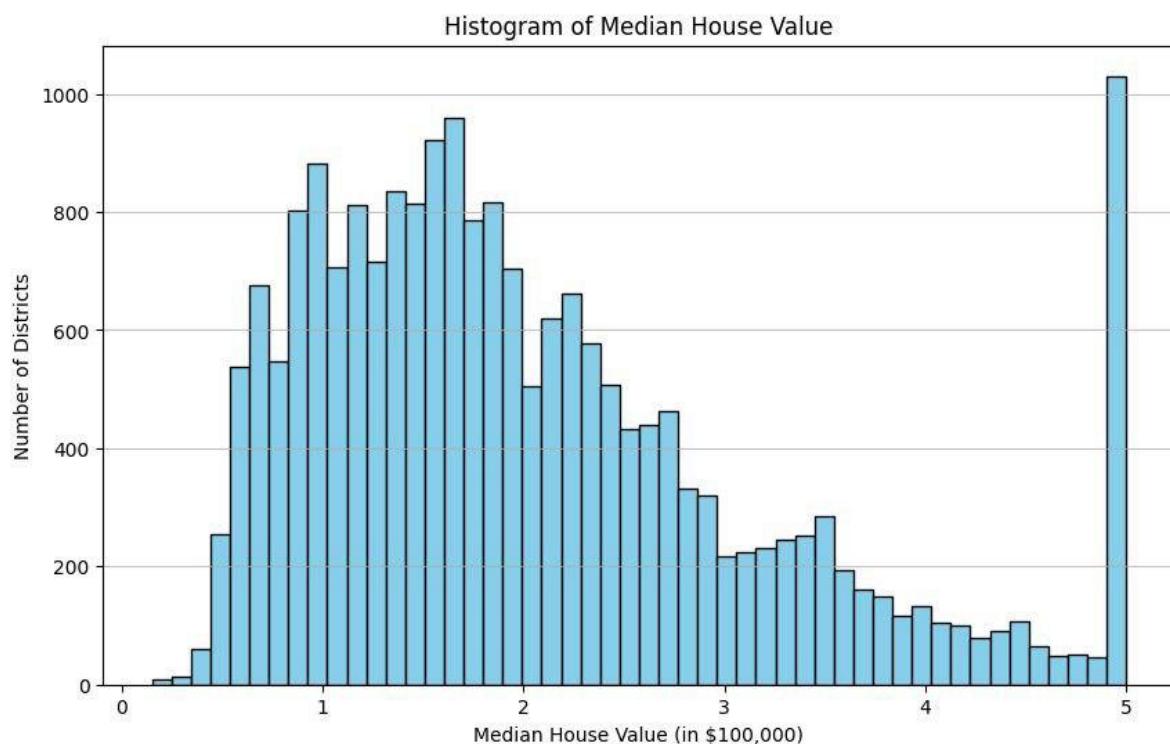


Figure 1: Histogram of Median House Value

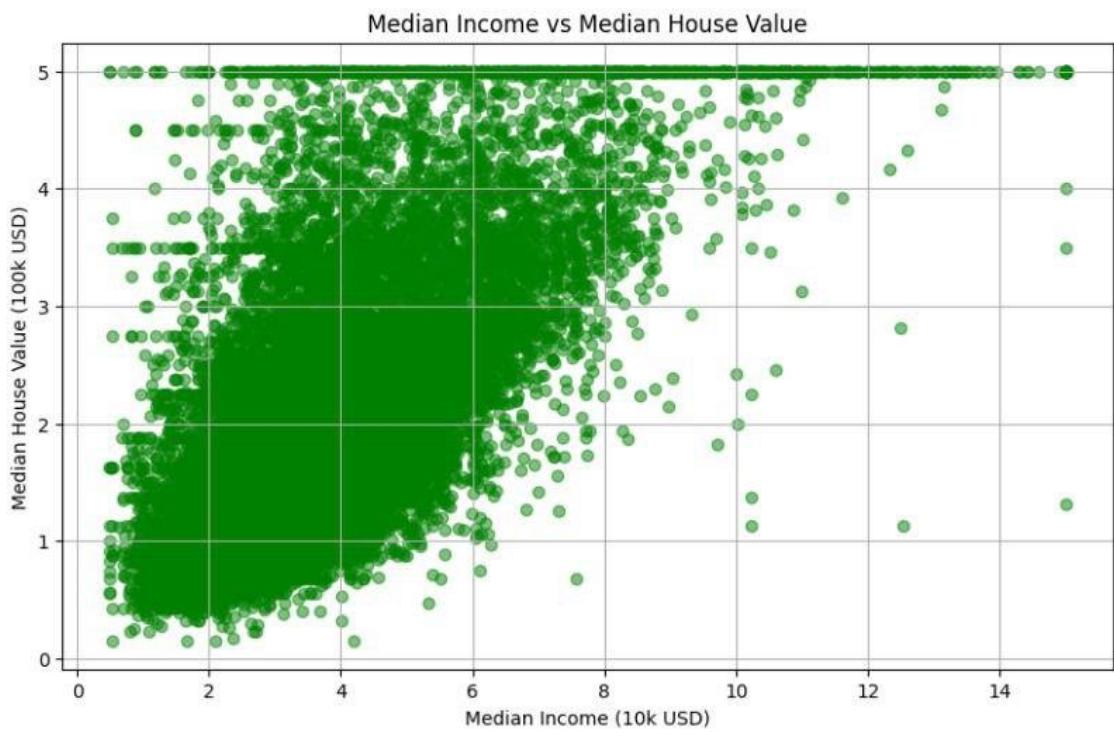


Figure 2: Median Income vs Median House Value

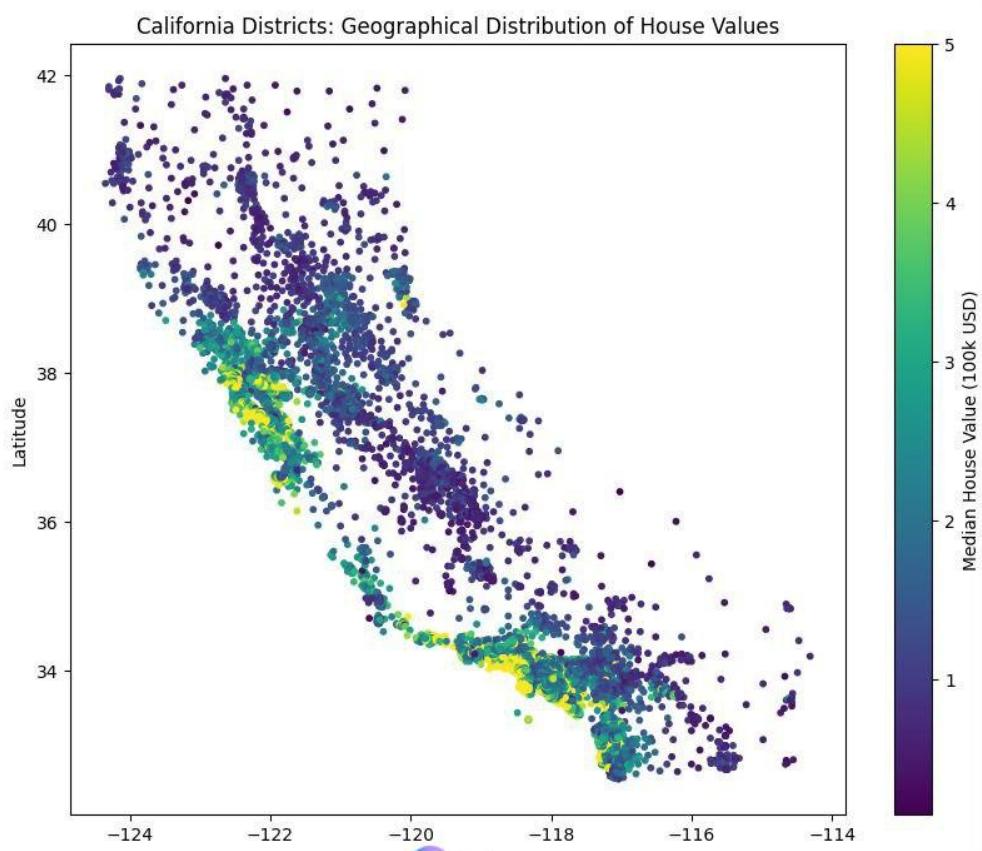


Figure 3: Geographical Distribution of house values

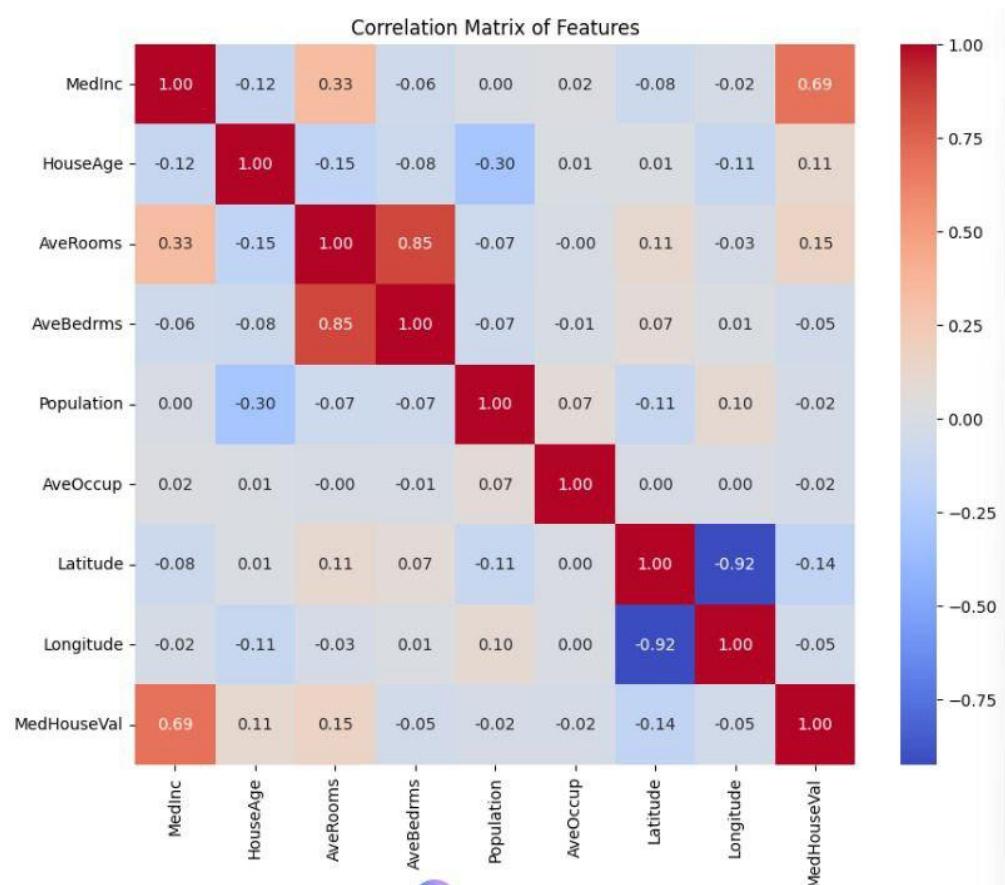


Figure 4: Correlation Matrix of features

4.3 Model performance

Table 3: Model Performance

Model	Mean Squared Error(MSE)	R ² Score	Interpritation
Linear Regression	0.53	0.64	Moderate predictive power, captures main trends
Decision Tree	0.25	0.78	Better fit; may overfit slightly but captures non-linear relations

4.4 Discussion

- Median Income was the most influential factor for predicting house value.
- Decision Tree Regressor performed better due to its ability to model complex, non-linear relationships.
- Linear Regression provided interpretability and smoother predictions but underperformed slightly in accuracy.

4. Conclusion

This analysis demonstrated how socioeconomic and geographical features affect housing prices in California.

Key findings include:

- Median income has the strongest impact on housing value.
- Decision Tree Regressor achieved higher accuracy ($R^2 \approx 0.78$).
- Linear Regression is simpler and provides reasonable interpretability.

Limitations:

- Dataset is from 1990; it does not reflect recent market trends.
- No categorical features like proximity to city centers or employment rate.

Applications:

- Useful for urban planning, real estate prediction, and housing policy formulation.
- Models can be extended using modern data for AI-driven real estate insights.

5. References

- [colab.ws](#)
- [Stack Overflow](#)
- [TutorialsPoint:\(Big Data Analytics - Charts & Graphs\)](#)
