Lab Project Report

Summer 2025

Course Code: CSE422
Course Title: Artificial Intelligence

Table of Contents

Introduction

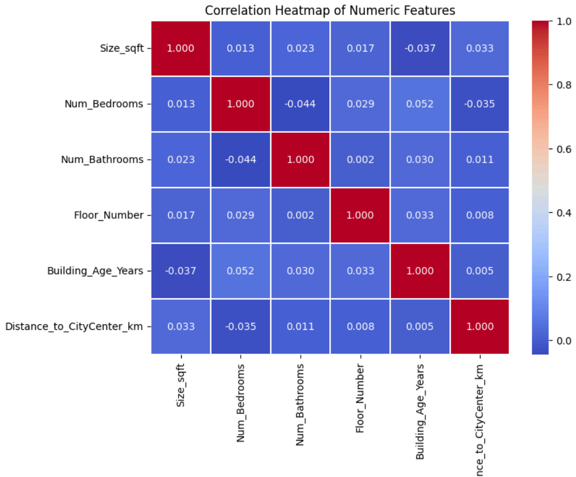
The aim of this project is to apply Machine Learning techniques to predict and analyze flat prices based on various features.

By training and testing different machine learning models, this project attempts to identify the most suitable approach for predicting housing price by doing comparison analysis of all the models.

The motivation behind this work is to provide a proper solution that helps buyers, sellers, and real estate agents make informed decisions on house pricing. With the regular change of house price and inflation, such predictive systems can save time and enhance the efficiency of the housing market.

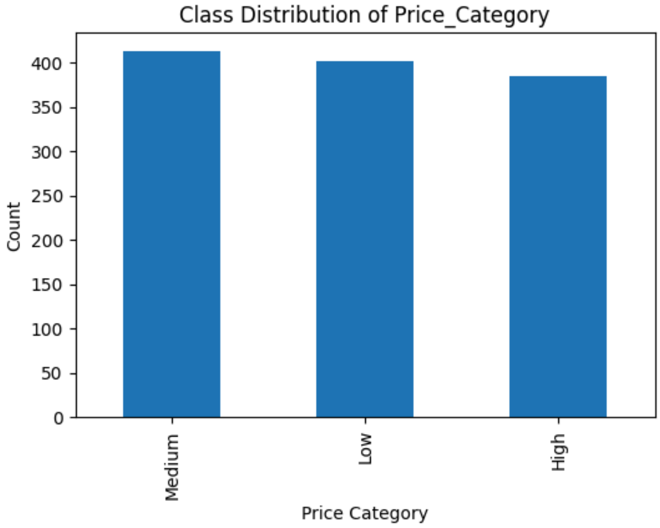
Dataset Description

- There are 12 features in the dataset where "Price_Category" is the target variable.
- It's a classification problem because the target variable has categorical values.
- There are 1200 datapoints in the dataset.
- There are both categorical and quantitative features in the dataset.
    - Categorical Features: Location, Has_Balcony, Parking_Available, Nearby_Schools, Security_Level, Price_Category.
    - Quantitative Features: Size_sqft, Num_Bedrooms, Num_Bathrooms, Floor_Number, Building_Age_Years, Distance_to_CityCenter_km.
- Yes, we need to encode the categorical features. Because, the machine learning algorithms cannot directly work with string values and require numerical values to work with.
- Correlation Heatmap:

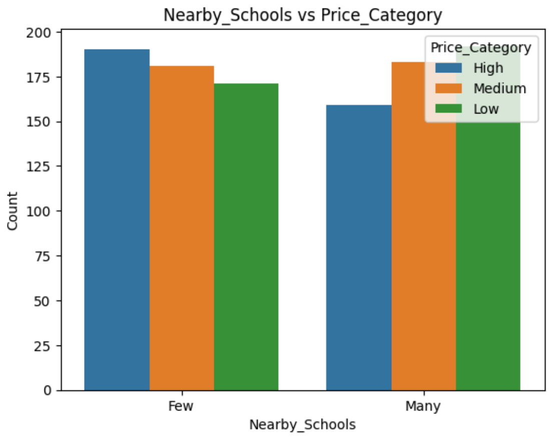


- From the correlation test, we see that the numerical features in the dataset are weakly correlated with each other. Which means each feature contributes to the target independently.

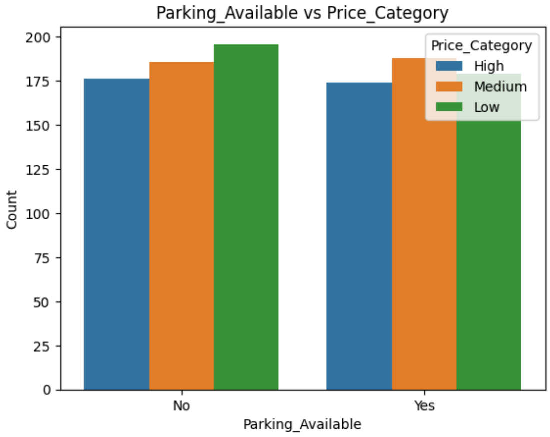
- No, in the output feature, all unique classes do not have an equal number of instances. Medium price flats had slightly more data points compared to the High and Low categories. However, the dataset is balanced enough to run machine learning algorithms.



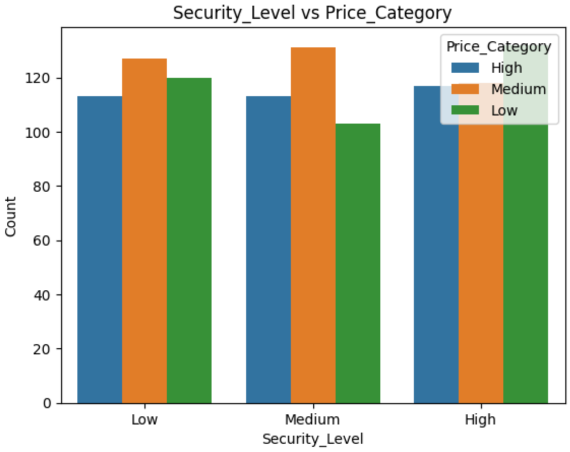
- Some important relationships extracted using Exploratory Data Analysis (EDA):
    - Flats in the Low price category generally have many schools nearby.



    - Flats in the Low price category generally have no parking availability.



- Flats in the Medium price category generally have low and medium security level.



Dataset pre-processing

- **Handling Null Values**
  - **Fault:** Machine learning models cannot directly process Null values also its not possible to encode or scale the dataset while Null value exists.
  - **Solution:** We used the simple imputer method instead of deleting the rows with null values because there are more than 143 rows with null values. Removing all of them will significantly reduce the datapoints. Also, removing the columns would aslo reduce the number of features significantly.
    - Numerical features with null values are replaced with the mean of the column.
    - Categorical features with null values are replaced with the mode of the column.
- **Encoding**
  - **Fault:** Machine learning models can only work with numbers, not strings.
  - **Solution:**

    - For data without hierarchy, we used the One Hot method to let the algorithm automatically split the "Location" column into 3 different columns Location_City Center, Location_Outskirts, Location_Suburbs.
    - For the rest of the categorical features, which are data with hierarchy, we used .map() to encode as the feature requires.

- **Feature Scaling**
  - **Fault:** Machine learning models will generally think features with bigger numbers are more important, which creates bias and inconsistency in result.
  - **Solution:** We applied Standard Scaler to all numerical features, which converts the data around the mean (0) and standard deviation (1).

Dataset Splitting

We used 70% of the data for training and 30% for testing, and the split was done without random shuffling to keep consistency. Also, used stratification so that all classes (Low, Medium, High) are balanced in both sets.

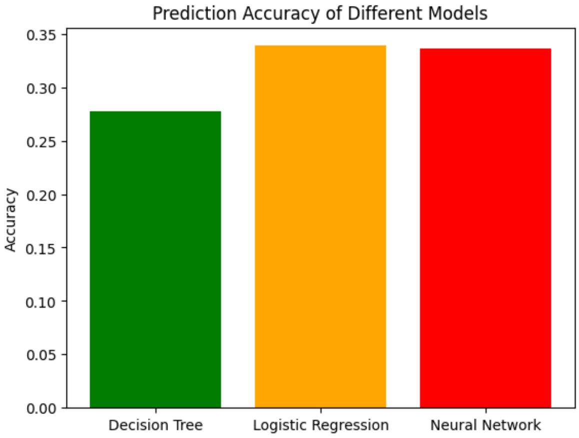
Model Training

As we have a classification problem, for supervised model training, we used:

1. Decision Tree
2. Logistic Regression
3. Neural Network

Comparision Analysis

- **Prediction Accuracy**

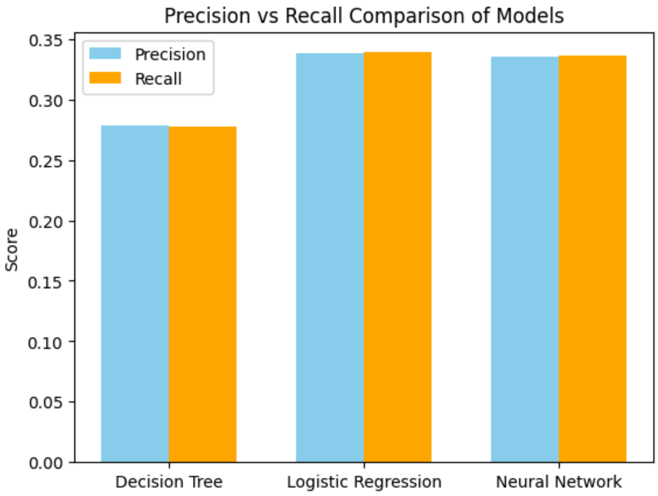


    - Decision Tree: 27.8%
    - Logistic Regression: 33.9%
    - Neural Network: 33.6%

Here, Logistic Regression performed slightly better than the others, and the Decision Tree had the lowest prediction accuracy.

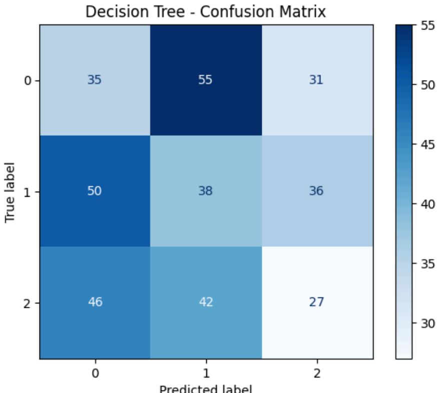
- **Precision & Recall**
    - Decision Tree
        - Precision: 27.9%
        - Recall: 27.8%
    - Logistic Regression
        - Precision: 33.9%
        - Recall: 33.9%

    - Neural Network
        - Precision: 33.5%
        - Recall: 33.6%

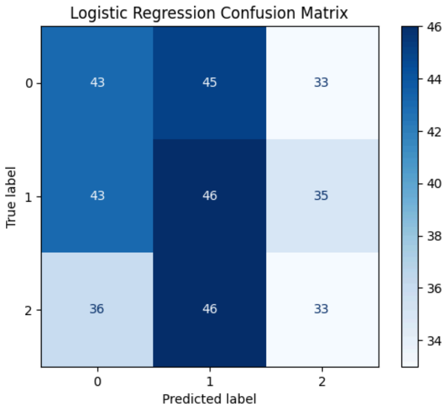


Here, Logistic Regression and Neural Network are quite close, while the Decision Tree is lacking.

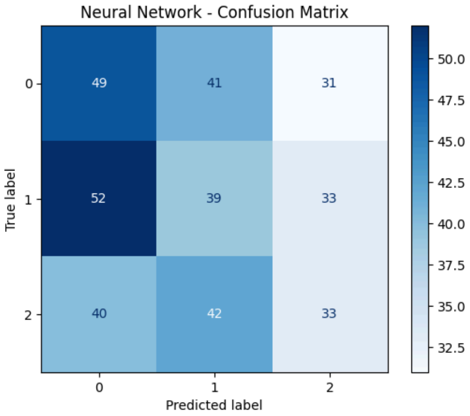
- **Confusion Matrix**



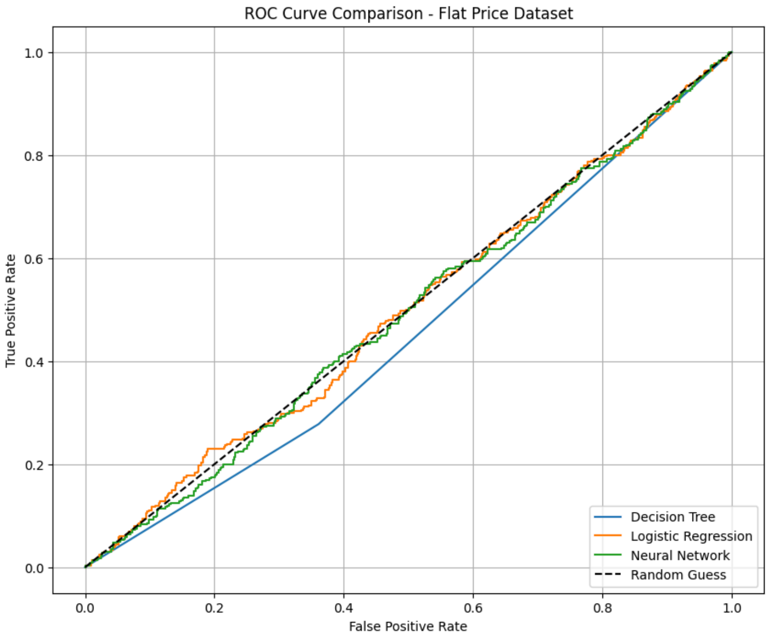
- Decision Tree: Very low diagonal values, much higher misclassifications.
- Logistic Regression: More balanced, correctly predicted low and medium more than high.



- Neural Network: Best predictions on low, but worse predictions for medium and high.



Overall, all models are confusing categories, but Logistic regression and Neural Network are a bit more stable.



- **ROC Curve & AUC**
    - Decision Tree: 0.457
    - Logistic Regression: 0.494
    - Neural Network: 0.494

Since all AUC scores are very close to 0.5, the models are just randomly guessing.

Conclusion

From the results of this project, it can be observed that different machine learning models perform differently when predicting flat prices. For the given dataset, all models are just randomly guessing. Thus, none of the models is a good choice to use.

The evaluation metrics, such as accuracy, precision, recall, and confusion matrices, highlighted that no single model is perfect for all scenarios. The results suggest that the features in the dataset do contain enough information to predict flat price categories, but some overlap and noise in the data can affect classification performance.

The performance of the models can be explained by the nature of the dataset. Some features do not have strong correlations with the target variable. Models like Neural Networks are more powerful but need more data to achieve better results. Logistic Regression assumes linear boundaries, which can give bad results with inconsistent data.

During the project, the main challenge was with the inconsistent dataset. With the dataset given, predicting with the ML models ar very challenging.