Predictive Modelling and Clustering for Housing Price Analysis

Course - Machine Learning – 1: Introduction (BUAN302B)

Name – Raghav Kajaria

Student ID – 220281

# 1. Introduction

The objective of this analysis is to use regression models to predict housing prices and unsupervised learning techniques to find significant clusters of housing features. The dataset includes important characteristics that affect home prices, such as the number of bedrooms, area, and parking. We seek to capture the relationship between housing features and price while identifying the most significant predictors by putting regression models like Linear Regression, K-Nearest Neighbours (KNN), Decision Tree Regressor, and Random Forest Regressor into practice.

Furthermore, clustering techniques like K-Means and Hierarchical Clustering assist in identifying organic groupings in the dataset, which can guide tactics like pricing tiers or market segmentation.   
The results offer a thorough examination of model performance, stressing both generalisability and accuracy, as well as information on the advantages and disadvantages of each strategy.

# 2. Data Processing and EDA

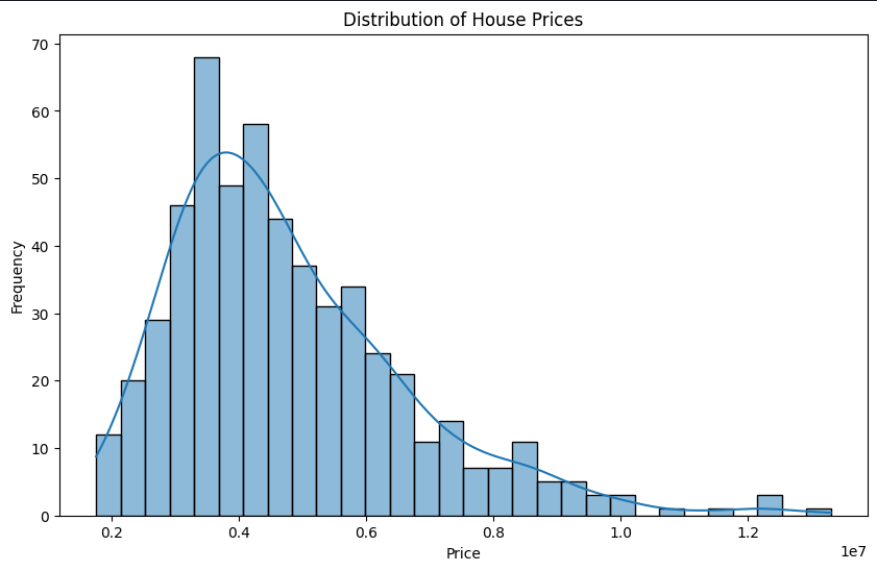
The dataset comprises 545 rows and 13 features, with the target variable being price. It was confirmed that no missing values were present, ensuring a clean starting point for analysis.

## Data Preprocessing

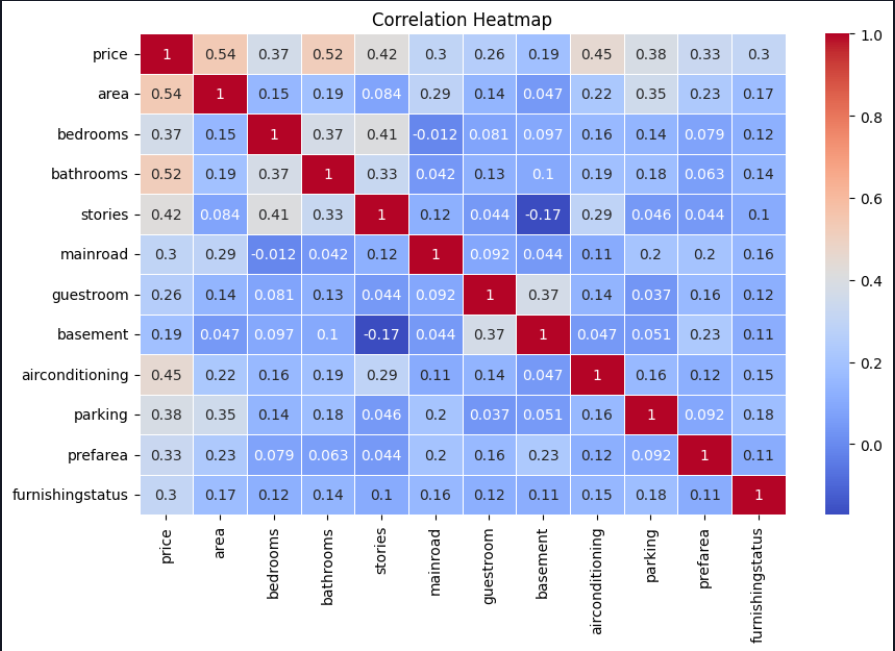
1. Transformation of Binary Variables: Features such as mainroad, guestroom, basement, and airconditioning were originally binary categorical variables (yes/no) and were transformed into numerical format, where yes = 1 and no = 0.
2. Encoding Categorical Features: The furnishingstatus feature was ordinally encoded with scores of 10 (furnished), 5 (semi-furnished), and 0 (unfurnished).
3. Scaling: Numerical features like area, bedrooms, bathrooms, stories, and parking were standardized using StandardScaler to ensure all variables contributed equally during model training.
4. Dimensionality Reduction: PCA was applied to retain 95% of the variance, reducing feature dimensions while maintaining critical information.

## Exploratory Data Analysis (EDA)

* Target Variable Distribution: The distribution of house prices showed a positive skew, with most houses priced between ₹2,000,000 and ₹6,000,000. High-priced houses were relatively sparse.



* Correlation Heatmap: Strong positive correlations were observed between price and features such as:
  + Area (0.54) – Larger homes are associated with higher prices.
  + Bathrooms (0.52) – Houses with more bathrooms tend to have higher prices.
  + Airconditioning (0.45) – Indicates that modern amenities significantly influence pricing.



Other features like mainroad and parking showed moderate correlations, while basement and guestroom had lower correlations.

# 3. Comparative Analysis and Insights

## Supervised Learning Models

| Model | Mean Squared Error (MSE) | R² Score |
| --- | --- | --- |
| Linear Regression | 1.08 × 10¹² | 0.62 |
| KNN Regressor | 1.32 × 10¹² | 0.54 |
| Decision Tree Regressor | 1.95 × 10¹² | 0.32 |
| Random Forest Regressor | 1.29 × 10¹² | 0.55 |

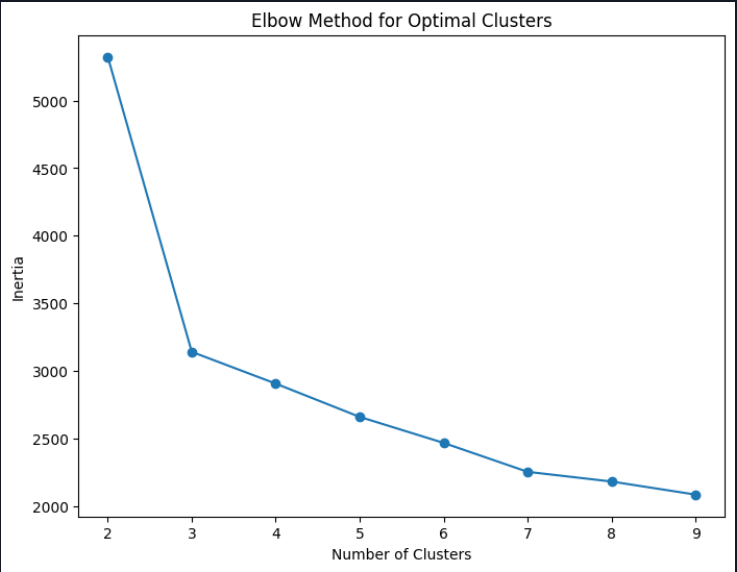
Among the supervised learning models, **Linear Regression** achieved the highest R² score of 0.62, indicating its strong ability to explain the variance in housing prices. However, its higher MSE reflects its sensitivity to extreme values, particularly high-priced houses. In contrast, **Random Forest** **Regressor** strikes a balance with a lower MSE (1.29 × 10¹²) and a reasonably good R² score (0.55). This makes Random Forest more robust for real-world scenarios, as it minimizes large prediction errors caused by outliers.

The **KNN Regressor** showed moderate performance, with an R² score of 0.54, but its higher MSE highlights its limitations in generalizing across varying data points. Lastly, the **Decision Tree Regressor** underperformed with the lowest R² score of 0.32 and the highest MSE (1.95 × 10¹²). Its shallow depth prevented it from effectively capturing complex relationships within the dataset.

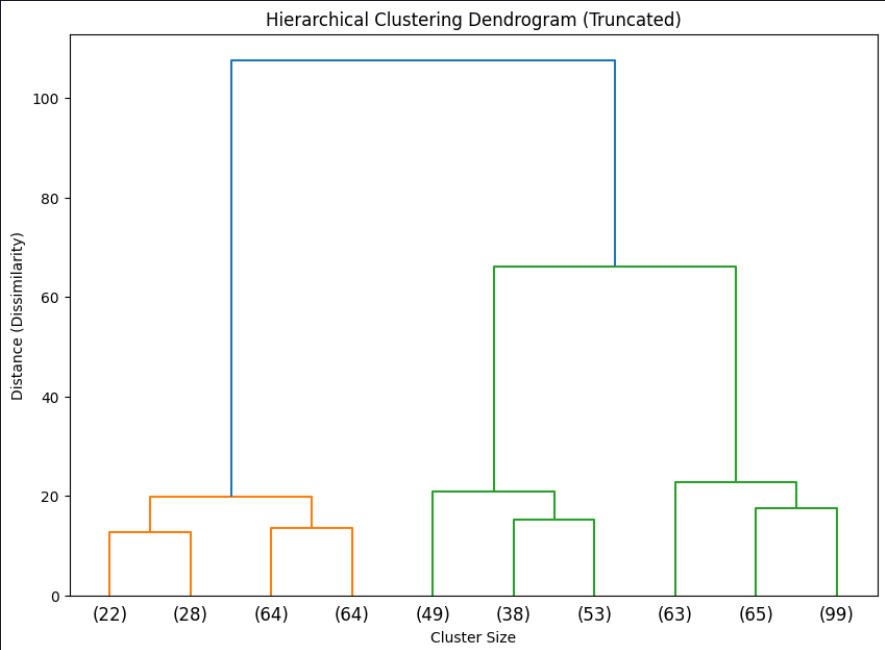
## Unsupervised Learning Models

| Clustering Model | Optimal Clusters | Silhouette Score |
| --- | --- | --- |
| K-Means Clustering | 3 | 0.45 |
| Hierarchical Clustering | 4 | 0.41 |

For unsupervised learning, **K-Means Clustering** achieved the best silhouette score of 0.45, indicating moderate cluster separation. The optimal number of clusters was determined to be three, where groups formed primarily based on key features like area and ‘furnishingstatus’. These clusters provide actionable segmentation insights, useful for identifying housing market segments.

**Hierarchical Clustering**, on the other hand, generated four clusters with a slightly lower silhouette score (0.41). While the dendrogram revealed a nested cluster structure, the lower score suggests greater overlap between groups. Hierarchical Clustering is useful for interpretability, as it provides a hierarchical view of group relationships, but it lacks the precision and clearer segmentation offered by K-Means.



# 4. Conclusion

Random Forest is the most reliable model for predicting housing prices, balancing accuracy and error minimisation, according to a comparative study of supervised learning models. Although Linear Regression has the highest R2 score, its practical applicability is limited due to its inability to handle outliers.

Compared to Hierarchical Clustering, which excels in interpretability but has overlapping clusters, K-Means Clustering in unsupervised learning offers more comprehensible and useful groupings.   
All things considered, this study illustrates the advantages and disadvantages of various machine learning models, offering insightful information for housing market segmentation and price prediction. To enhance performance, future research can concentrate on incorporating sophisticated models and additional feature engineering.

Predictive Modelling and Clustering for Mobile Price Range Analysis

Course - Machine Learning – 1: Introduction (BUAN302B)

Name – Raghav Kajaria

Student ID – 220281

# **1. Introduction**

This project's main goal is to develop predictive and clustering models for analysing and classifying mobile phones into various price ranges according to a variety of functional and technical features. The dataset includes a wide variety of attributes, such as camera specifications, screen resolution, RAM, and battery life. The project intends to discover possible clusters within the dataset for improved market segmentation and predict whether a phone belongs to a "low" or "high" price category by utilising both supervised and unsupervised learning techniques.

The classification task uses supervised models such as K-Nearest Neighbours (KNN), Random Forest, Decision Tree, Naive Bayes, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). For unsupervised learning, K-Means and Hierarchical Clustering are employed to uncover hidden patterns and groupings within the data. The outcomes of these analyses provide actionable insights for pricing strategies and feature prioritization.

# **2. Data Processing and EDA**

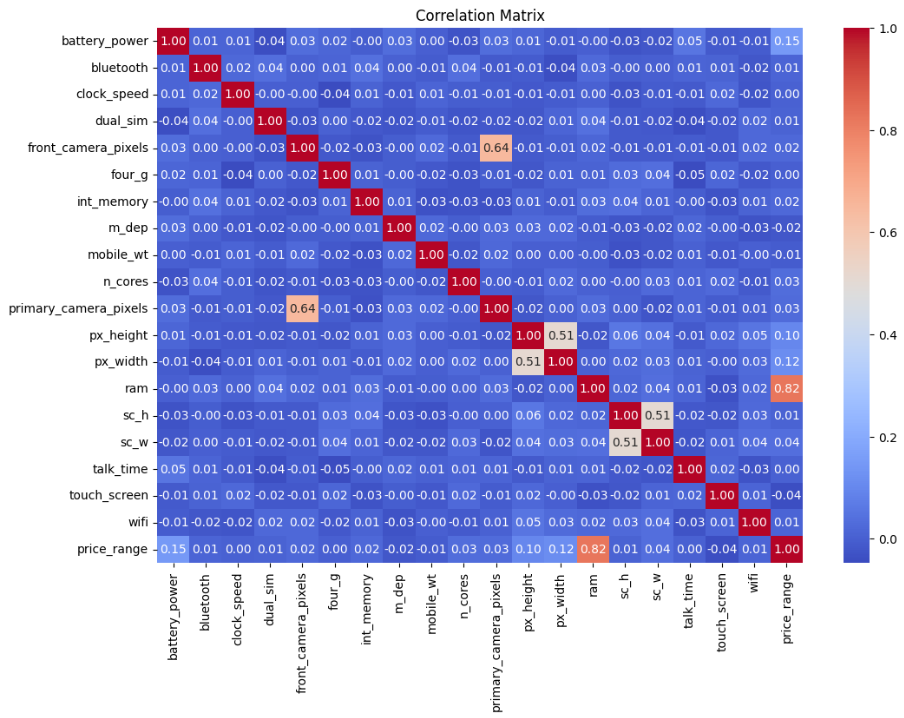
The dataset consists of 2000 observations with 20 features, including a binary target variable (price\_range) representing low (0) and high (1) price categories. Initially, the dataset was multi-class, the classes being low (0), moderate (1), high (2), very high (3), but for convenience, it was converted into a binary classification problem, with 0 representing low and 1 representing high price ranges. Data preprocessing included standardization of numerical features using StandardScaler and dimensionality reduction using Principal Component Analysis (PCA) for unsupervised learning. No missing values were present, ensuring a clean dataset for model implementation.

Exploratory Data Analysis (EDA) revealed critical insights. The target variable was balanced, ensuring that no class bias affected the classification models. Features such as ram, px\_height, and battery\_power exhibited strong correlations with the target variable, suggesting their importance in price prediction. On the other hand, binary features like wifi, bluetooth, and four\_g had minimal correlation, indicating limited predictive utility. Distributions of numerical features varied widely, with some, such as front\_camera\_pixels and primary\_camera\_pixels, showing significant skewness.



The correlation matrix highlighted the interdependencies among features, such as the moderate correlation between px\_height and px\_width. These relationships were considered during dimensionality reduction and model development to ensure optimal performance.

The correlation matrix highlighted the interdependencies among features, such as the moderate correlation between px\_height and px\_width. These relationships were considered during dimensionality reduction and model development to ensure optimal performance.



# **3. Comparative Analysis and Insights**

## **3.1 Supervised Learning**

The following table summarizes the performance metrics of the models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **AUC-ROC** |
| Random Forest | 0.97 | 0.97 | 0.97 | 0.97 | 1.00 |
| LDA | 0.97 | 0.97 | 0.97 | 0.97 | 1.00 |
| QDA | 0.96 | 0.96 | 0.96 | 0.96 | 1.00 |
| Decision Tree | 0.94 | 0.94 | 0.94 | 0.94 | 0.96 |
| Naive Bayes | 0.95 | 0.96 | 0.95 | 0.95 | 0.99 |
| KNN | 0.91 | 0.91 | 0.91 | 0.91 | 0.98 |

Insights:

* **Random Forest and LDA** emerged as the top-performing models with their perfect AUC-ROC scores and high accuracy, precision, recall, and F1 scores. These models are highly reliable for binary classification tasks and leverage strong feature interactions.
* **Naive Bayes and Decision Tree** provided good performance, with Naive Bayes excelling in computational efficiency and Decision Tree offering interpretable decision pathways.
* **KNN** had the lowest performance among the models due to its sensitivity to feature overlaps and noise. While effective for smaller datasets, it struggled with the complexity of this data.

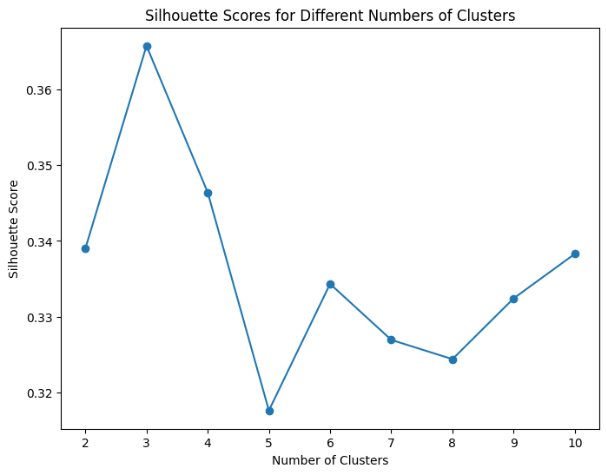
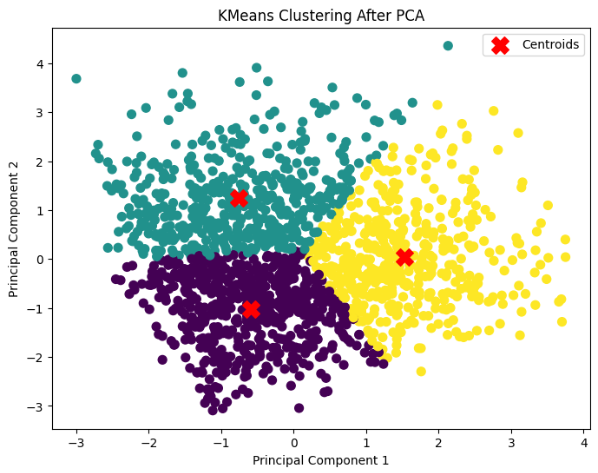
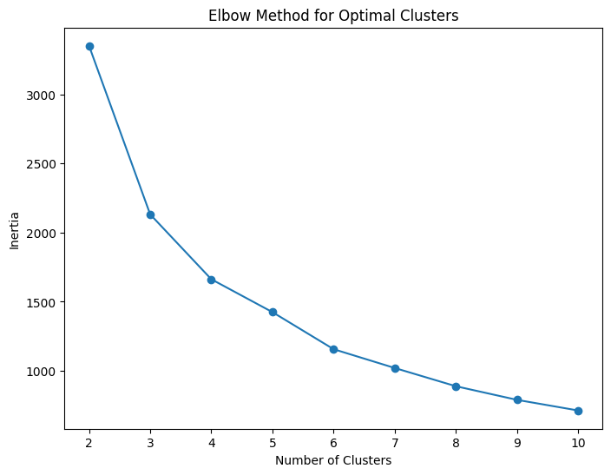
## **4.1 Unsupervised Learning**

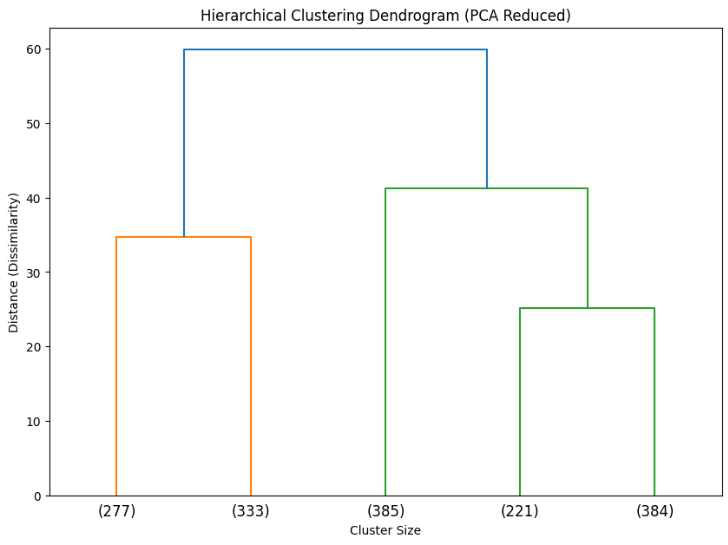
The following table summarizes the clustering metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Clustering Model** | **Optimal Clusters** | **Silhouette Score** | **Interpretability** |
| K-Means | 3 | 0.37 | Moderate cluster overlap |
| Hierarchical | 2 | 0.34 | Balanced hierarchical group |

Insights:

* **K-Means Clustering** showed moderate separation of clusters, as evident from the silhouette score and scatter plot. While the clusters provide useful segmentation insights, their overlap suggests some inherent feature similarities among groups.
* **Hierarchical Clustering** revealed a hierarchical structure with well-balanced clusters but slightly lower silhouette scores. This approach is beneficial for understanding nested groupings but may not provide clear boundaries for practical applications.



# **4. Conclusion**

This study demonstrated the effectiveness of machine learning in classifying mobile phones into price ranges. Random Forest and LDA stood out as the top performers, with near-perfect metrics and excellent class separability. EDA highlighted the importance of features like ram and battery\_power, which played pivotal roles across models. Unsupervised learning using K-Means and Hierarchical Clustering provided additional insights into data segmentation, albeit with moderate silhouette scores indicating overlapping clusters.

Future work could involve exploring advanced feature engineering techniques, addressing class overlaps in clustering, and testing models on larger, more diverse datasets to validate their generalizability. Overall, the findings provide valuable guidance for feature prioritization and predictive modelling in pricing strategies.