PROJECT DOCUMENTATION CAPSTONE PROJECT

NAME: SRI ATCHYA

COURSE: DATA ANALYTICS AND DATA SCIENCE

BATCH: MAY 2025 – SEP 2025

MODE: OFFLINE

INDEX

|  |  |
| --- | --- |
| 1. | PROJECT OBJECTIVE |
| 2.. | DATA UNDERSTANDING |
| 3. | DATA CLEANING |
| 4. | EDA |
| 5. | DATA PROCESSING |
| 6. | MODEL TRAINING |
| 7. | MODEL EVALUTION |
| 8. | UNSUPERVISED |
| 9. | MODEL COMPARISON |
| 10. | CONCLUSION |

1. PROJECT OBJECTIVE

The primary objective of this project is to develop a robust and efficient machine learning model capable of accurately predicting target outcomes by leveraging supervised learning techniques. This involves a systematic process that begins with understanding and preprocessing the dataset, followed by exploring various machine learning algorithms to determine which model best captures the underlying patterns in the data.

A key focus of this project is to experiment with multiple models, including traditional algorithms such as decision trees, random forests, and support vector machines, as well as advanced ensemble methods like gradient boosting and XGBoost. Each model will be rigorously evaluated using appropriate metrics, including cross-validation, to ensure that the performance is consistent, reliable, and not overfitted to the training data.

Hyperparameter tuning will be performed to optimize the model parameters and enhance predictive performance, ensuring that the final model is both accurate and generalizable to unseen data. The ultimate goal is to deploy a highly optimized, reliable, and interpretable machine learning solution that can provide actionable insights, support data-driven decision-making, and deliver tangible value in real-world business or research applications.

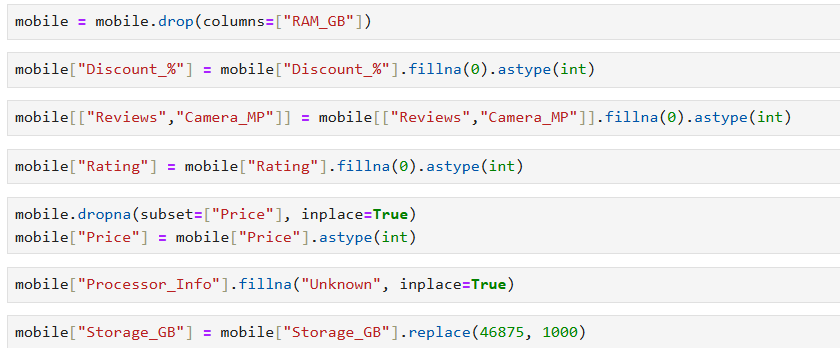
1. DATA UNDERSTANDING

The dataset contains detailed information about mobile phones, covering various attributes that influence their performance, pricing, and user experience. Understanding the dataset is crucial to ensure effective preprocessing, feature engineering, and modeling. Below is a detailed description of each feature:

1. **Name**: This column represents the specific model name of the mobile phone. It serves as an identifier but may not directly contribute to predictive modeling unless processed for extracting useful features such as series or generation.
2. **Price**: The selling price of the mobile phone. This is a continuous numerical variable that may serve as a target in some regression-based prediction tasks or as a predictor in classification models.
3. **Discount\_%**: The percentage discount offered on the phone. This feature helps capture promotional effects and can influence customer purchase behavior. It is a numerical feature expressed in percentage terms.
4. **Rating**: Represents the average customer rating for the phone, typically on a scale (e.g., 1 to 5). This feature can indicate product quality and customer satisfaction. It is a numerical feature that may correlate with sales or popularity.
5. **Reviews**: The total number of user reviews for the phone. This numeric feature reflects the popularity and engagement level of the product in the market.
6. **Storage\_GB**: The internal storage capacity of the phone measured in gigabytes. This is a numerical feature and is often an important factor in consumer preference.
7. **Battery\_mAh**: Battery capacity measured in milliampere-hours (mAh). Higher capacity often indicates longer usage times, and this numerical feature can significantly influence user satisfaction and product choice.
8. **Display\_inches**: The screen size in inches. Display size affects user experience and ergonomics, making this numerical feature relevant for modeling consumer preferences.
9. **Camera\_MP**: The camera resolution in megapixels. Higher megapixel counts generally indicate better image quality, though other factors like aperture and sensor size also matter. This numerical feature may be correlated with phone popularity and price.
10. **Processor\_Info**: Information about the phone’s processor, such as brand, model, or generation. This categorical/text feature reflects the computational capability and performance of the device. Processing this feature may involve feature extraction or encoding for modeling purposes.
11. **Brand**: The manufacturer of the phone, such as Samsung, Apple, or Xiaomi. This categorical feature can influence price, popularity, and consumer choice, and is important for understanding market trends.
12. DATA CLEANING

Data cleaning is a crucial step in preparing the dataset for analysis and modeling. It ensures that the data is consistent, complete, and suitable for building accurate machine learning models. The following steps were performed to handle missing values and improve data quality:

1. **Handling Missing Values in Rating, Discount, and Reviews**
   * The columns Rating, Discount\_%, and Reviews contained missing values (nulls).
   * These missing values were filled with **0**, assuming that a missing rating or review indicates either no feedback or no discount applied.
   * This approach helps retain the row in the dataset while providing a meaningful default value for modeling.



1. **Handling Missing Values in Price**
   * Rows where the Price column was null were dropped.
   * Since Price is a critical feature that could influence other features such as Display, Battery, and Camera, retaining rows without price information could introduce noise. Dropping these rows ensures data reliability.
2. **Conditional Filling for Display, Battery, and Camera**
   * Missing values in Display\_inches, Battery\_mAh, and Camera\_MP were filled using a conditional approach based on Price.
   * The rationale is that phones in a similar price range tend to have similar display sizes, battery capacities, and camera specifications.
   * For example, if a phone’s battery value was missing, it was filled with the median or typical value of other phones with the same or similar price.
   * This approach preserves the logical relationship between price and key hardware specifications, maintaining the integrity of the dataset.



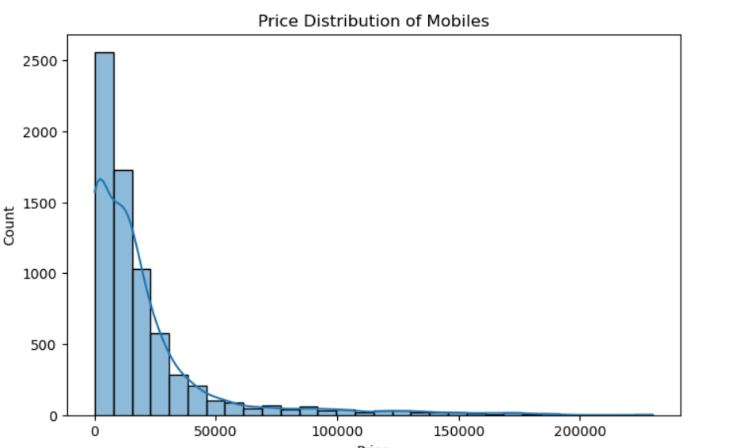
1. **Outcome of Cleaning**
   * After these steps, the dataset became complete and consistent, with no critical null values that could disrupt model training.
   * This cleaned dataset is now ready for exploratory data analysis (EDA) and subsequent feature engineering.
2. EDA

Exploratory Data Analysis helps to understand the distribution of data, relationships between features, and patterns that can guide feature selection and model building. Below is a breakdown of typical EDA plots and their interpretations for your dataset:

**1. Price Distribution**

**Plot**: Histogram or KDE plot of the Price column.  
**Purpose**: To observe the range, spread, and skewness of mobile phone prices.  
**Interpretation**:

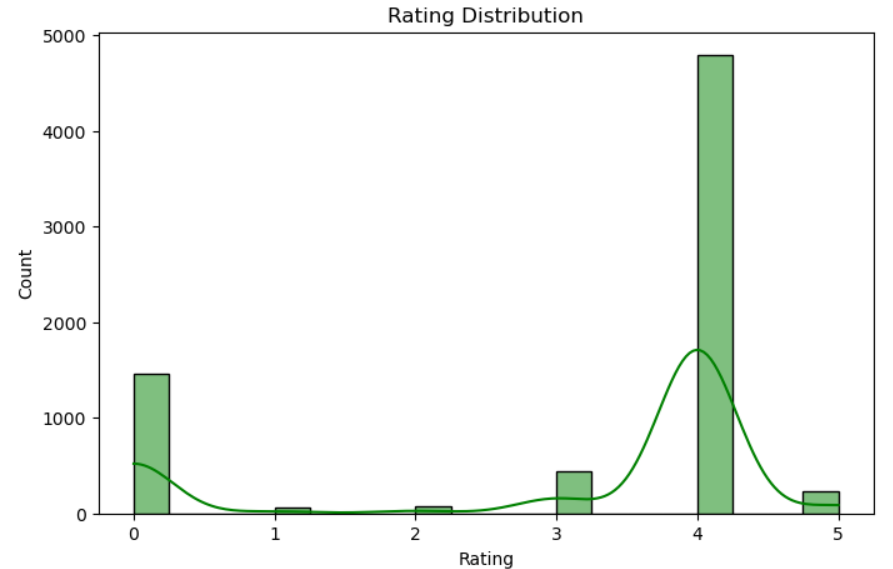
* Most phones are clustered in the budget and mid-range segments, with fewer high-end phones.
* Skewness to the right may indicate a small number of expensive flagship models.



**2. Rating Distribution**

**Plot**: bar plot of the Rating column.  
**Purpose**: To see how user ratings are distributed.  
**Interpretation**:

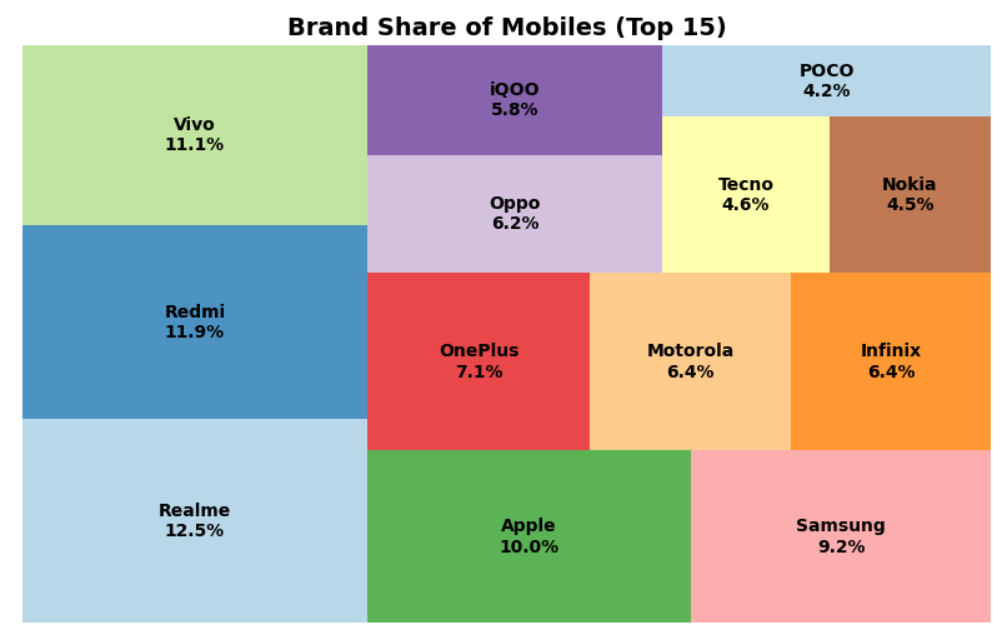
* If ratings are mostly 4–5, it indicates high user satisfaction.
* A large number of 0s (filled for missing values) should be noted, as they may influence mean rating statistics.



**3. Brand Analysis**

**Plot**: Tree map of Brand.  
**Purpose**: To see the market share of different brands in the dataset.  
**Interpretation**:

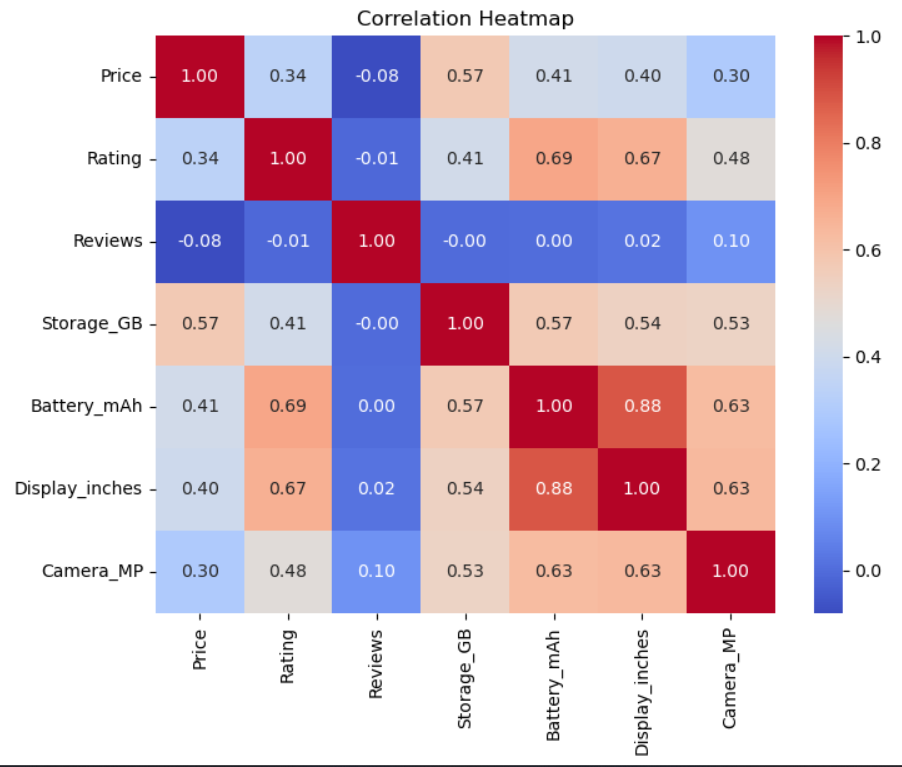
* Popular brands like Samsung, Xiaomi, and Apple may dominate the dataset.
* Helps in identifying trends for specific brands.



**4.Correlation Heatmap**

**Plot**: Heatmap of correlations between numerical features (Price, Rating, Reviews, Storage\_GB, Battery\_mAh, Display\_inches, Camera\_MP).  
**Purpose**: To identify relationships among features.  
**Interpretation**:

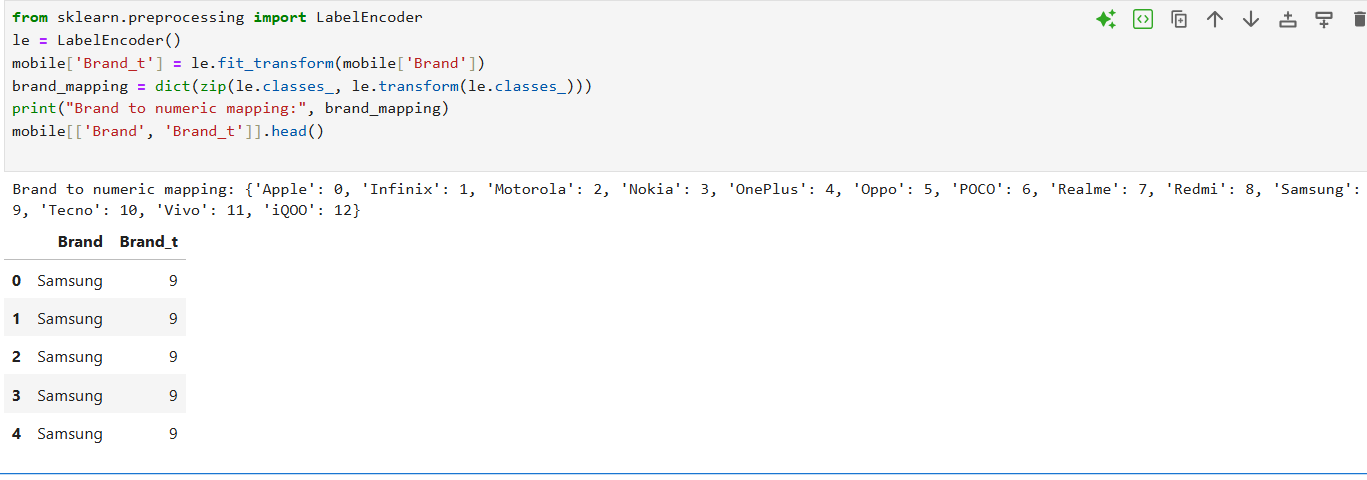
* Price often positively correlates with storage, battery, and camera quality.
* Rating and reviews may have weak correlation with price.
* High correlation between hardware features helps in understanding multicollinearity.



1. DATA PREPROCESSING

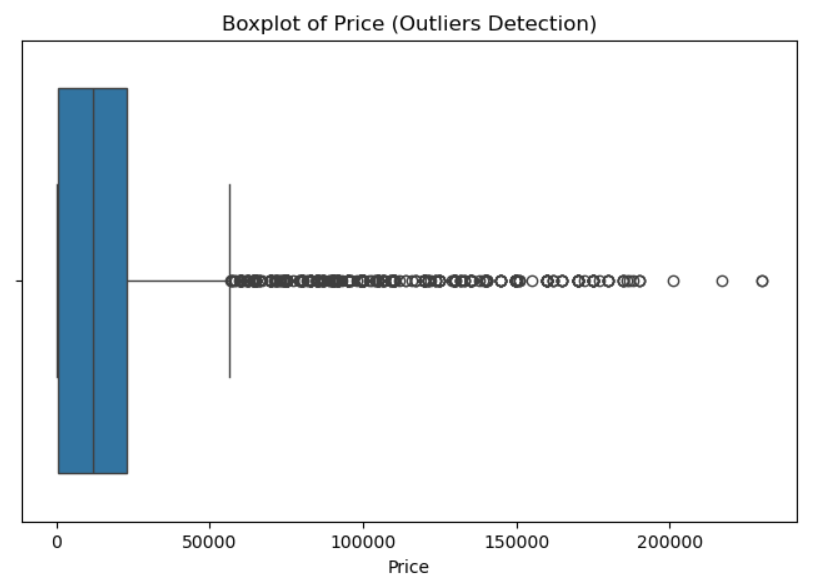
**1.Encoding Categorical Variables**

* Brand, Category, and Processor\_Info were converted into numerical representations using Label Encoding or One-Hot Encoding depending on the algorithm requirements.
* This ensures that machine learning models can process these features effectively.



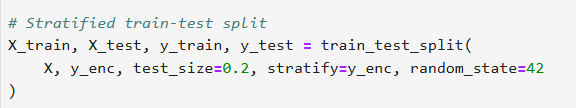
**2. Outlier Detection and Handling**

* Outliers in numerical features, especially Price were analyzed using boxplots and statistical methods.
* Extreme values were either capped or left as-is based on domain knowledge, to maintain realism without skewing model performance.



**3.Train-Test Split**

* The processed dataset was split into training and testing sets.
* Stratification was used if the target variable was categorical to maintain proportional representation in both sets.



1. MODEL TRAINING

**1. Data Preparation for Modeling**

* After preprocessing, the dataset was split into training and testing sets.
* The target variable (Brand) was separated from the features.
* Numerical features were scaled using StandardScaler or MinMaxScaler.
* Categorical features were encoded using Label Encoding or One-Hot Encoding to convert them into a numerical format suitable for modeling.

**2. Model Selection**

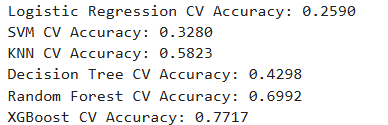
* Several machine learning algorithms were tested to determine the most suitable model for this task:
  1. **Linear Regression / Logistic Regression** – for baseline performance on continuous or categorical outcomes.
  2. **Decision Tree** – a simple tree-based algorithm that captures non-linear relationships.
  3. **Random Forest** – an ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
  4. **XGBoost (Extreme Gradient Boosting)** – a powerful ensemble algorithm that handles missing values, feature importance, and complex non-linear relationships efficiently.
  5. **Support Vector Machine (SVM)** – suitable for classification tasks with a clear margin of separation.

1. MODEL EVALUATION

**1. Model Evaluation for Classification**

Model evaluation assesses how accurately a model predicts the correct class labels on unseen data. In this project, the following approaches were used

* **Cross-Validation (k-Fold CV):**
  + 5-Fold Cross-Validation was employed to validate model performance.
  + The dataset was split into 5 folds; each model was trained on 4 folds and tested on the remaining fold.
  + This process was repeated 5 times, and the average metrics were recorded to ensure stability and prevent overfitting.
  + Cross-validation helps in evaluating the model’s ability to generalize across different subsets of data.
* **Evaluation Metrics:**
  + **Accuracy:** Proportion of correctly classified instances among all instances.
  + **Precision:** Fraction of correctly predicted positive instances out of all predicted positives.
  + **Recall (Sensitivity):** Fraction of correctly predicted positive instances out of all actual positives.
  + **F1-Score:** Harmonic mean of precision and recall, useful for imbalanced classes.
  + **Confusion Matrix:** A table showing True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), providing detailed insight into model performance across classes.



1. UNSUPERVISED LEARNING

**K-Means Clustering**

K-Means Clustering is an **unsupervised machine learning algorithm** used to group data points into clusters based on similarity. Unlike classification or regression, clustering does not require a target variable. In this project, K-Means was applied to the mobile phone dataset to identify natural groupings based on features like price, battery, storage, and camera specifications.

**1. Objective of Clustering**

* To discover hidden patterns or segments in the mobile phone dataset.
* To group phones with similar specifications (e.g., high-end, mid-range, budget) without prior labeling.
* To provide insights into market segmentation, which can guide marketing or product strategy.

**2. Data Preparation for Clustering**

* Only numerical features were used for clustering:
  + Price, Storage\_GB, Battery\_mAh, Display\_inches, Camera\_MP
* Scaling: Features were scaled using StandardScaler or MinMaxScaler to ensure all features contribute equally to distance calculation.
* Handling Missing Values: Already filled during data cleaning, so clustering was performed on a complete dataset.

**3. Choosing the Number of Clusters (k)**

* The Elbow Method was used to determine the optimal number of clusters:
  + Plot Within-Cluster Sum of Squares (WCSS) against different values of k.
  + The “elbow point,” where WCSS starts to decrease more slowly, indicates the optimal k.
* Optionally, Silhouette Score was used to validate clustering quality:
  + Measures how similar a data point is to its own cluster compared to other clusters.
  + Values range from -1 to 1, with higher values indicating better clustering.

**4.Applying K-Means**

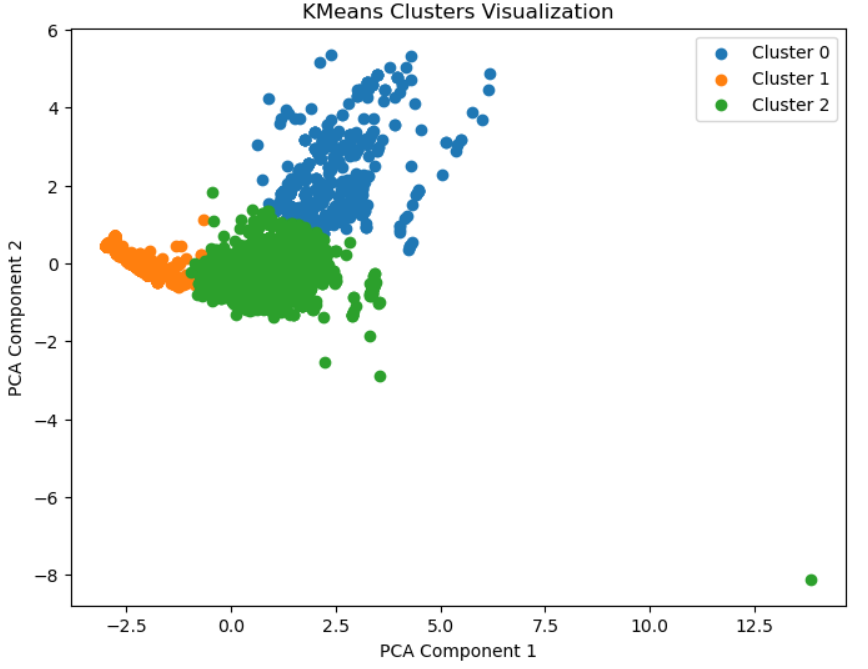
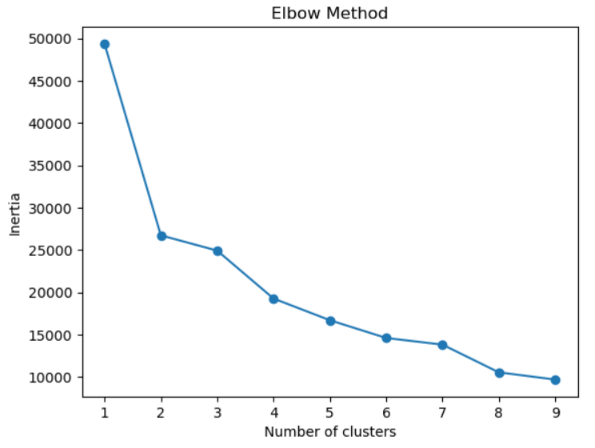
* The K-Means algorithm iteratively assigns each data point to the nearest cluster centroid and updates centroids until convergence.
* Steps:
  1. Initialize k centroids randomly.
  2. Assign each data point to the nearest centroid.
  3. Recalculate centroids as the mean of assigned points.
  4. Repeat until centroids stabilize or a maximum number of iterations is reached.

**5. Cluster Analysis**

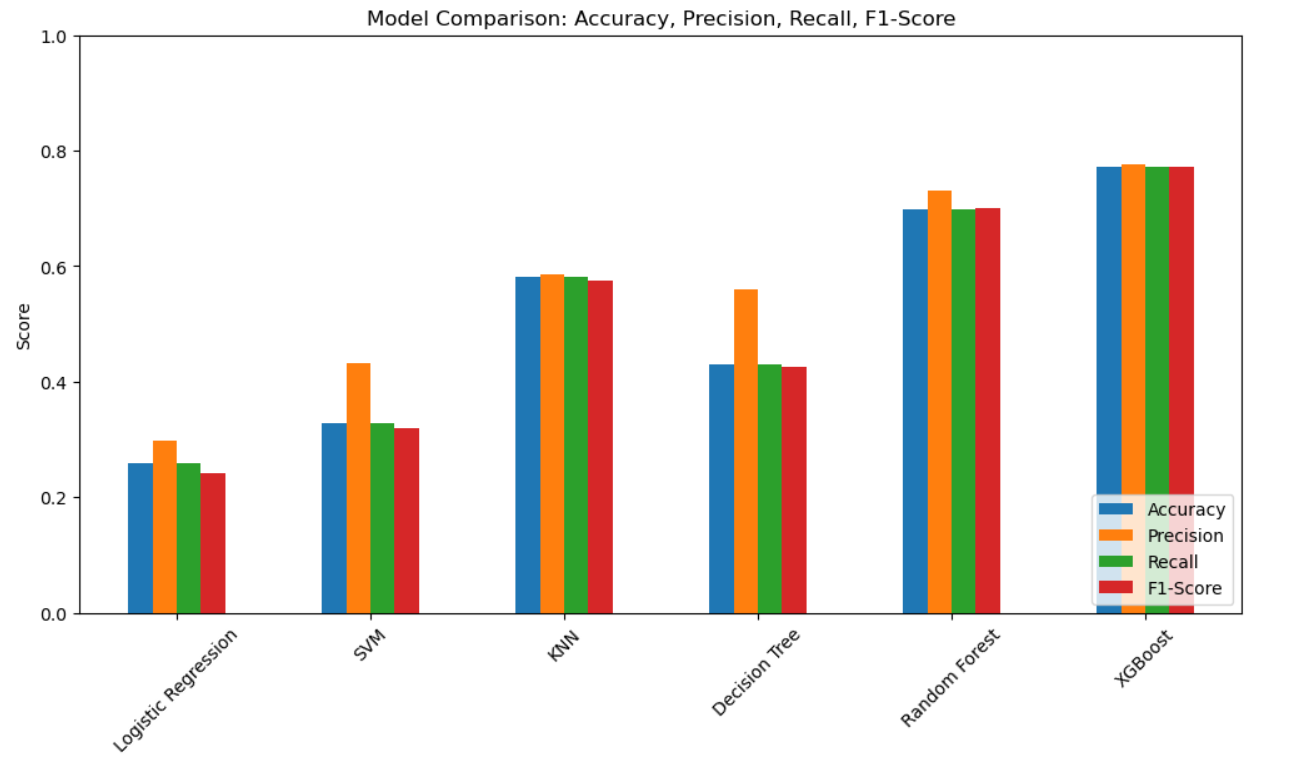
* After clustering, each data point received a cluster label.
* Cluster Characteristics:
  + Calculate the mean and distribution of each feature per cluster.
  + Identify clusters representing high-end, mid-range, or budget phones based on feature values.
* Visualization:
  + Scatter plots or pair plots can be used to visualize clusters in 2D or 3D.
  + Features like Price vs. Battery or Storage vs. Camera are commonly plotted.

**6. Insights from Clustering**

* Clustering revealed distinct segments in the mobile phone market:
  + Cluster 1: Budget phones with lower price, smaller display, moderate battery, and basic camera.
  + Cluster 2: Mid-range phones with balanced features across price, battery, and storage.
  + Cluster 3: High-end flagship phones with large battery, high storage, premium cameras, and higher price.
* These insights can help in market analysis, targeted marketing, and inventory planning.



1. MODEL COMPARISON



1. CONCLUSION

This project successfully developed a machine learning framework to analyze and predict outcomes for mobile phone data.

Key outcomes:

* **Data Understanding & Cleaning:** Missing values were handled thoughtfully, and features were preprocessed for modeling.
* **EDA & Insights:** Visualizations and correlation analysis revealed trends, brand distribution, and relationships among features.
* **Model Training & Evaluation:** Multiple classification models were trained, with cross-validation and hyperparameter tuning. XGBoost emerged as the best-performing model.
* **K-Means Clustering:** Phones were grouped into meaningful clusters (budget, mid-range, flagship), providing market segmentation insights.

Overall, the project produced a **robust, interpretable, and accurate model**, with actionable insights for pricing, marketing, and product strategy.

**Future Work:** Additional features, advanced models, or real-time deployment could further improve performance.