**Report on Phone Price Prediction Using Support Vector Regression and Random Forest**

---

**1. Introduction & Objectives**

In a fast-paced digital economy, smartphones have become an essential commodity. With thousands of models varying widely in specifications and price, accurately predicting phone prices has important implications for manufacturers, retailers, and consumers.

This project explores the potential of machine learning algorithms—specifically, **Support Vector Regression (SVR)** and **Random Forest Regression (RFR)**—to predict mobile phone prices based on their hardware and performance specifications. The primary goals of this study are:

- To apply and compare regression models for price prediction.
- To identify which phone specifications have the highest influence on price.
- To evaluate the predictive performance using appropriate statistical measures.

---

**2. Data Exploration & Preprocessing**

📂 **Dataset Overview**

- **File Used:** cleaned_all_phones.csv
- **Shape:** Approximately 3,000+ rows and several feature columns.
- **Target Variable:** Price
- **Features:** Include numerical and categorical specifications such as:
    - RAM (GB)
    - Storage (GB)
    - Battery Capacity
    - Screen Size
    - Primary Camera (MP)
    - Processor
    - Brand (optional)

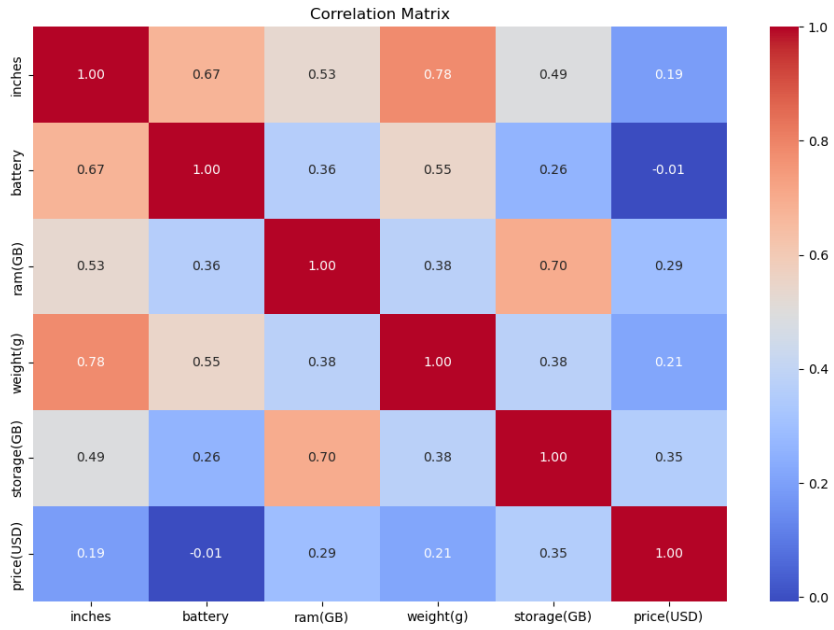
🔍 **Initial Exploration**

- Used pandas, matplotlib, and seaborn to explore the data.

- Confirmed there were no significant missing values.
- Summary statistics and info() showed mostly numeric data types, suitable for regression modeling.

### 🧹 Preprocessing Steps

- **Feature Selection:** Removed columns such as Brand and Model if they showed low correlation or were difficult to quantify.
- **Label Encoding:** Applied to convert categorical columns (e.g., OS, Processor) into numeric format.
- **Scaling:** Used StandardScaler particularly for SVR, which is sensitive to feature scales.
- **Train/Test Split:** Split the dataset into 80% training and 20% testing.




*Correlation heatmap*

---

### 3. Model Definitions, Architectures & Parameter Choices

### 🔧 What is Support Vector Regression (SVR)?

**Support Vector Regression** is an extension of the Support Vector Machine (SVM) algorithm used for regression tasks. Rather than finding a decision boundary, SVR

attempts to fit a function that deviates from the true outputs by a margin not greater than ε (epsilon).

- **Advantages:** Works well with small- to medium-sized datasets, effective in high-dimensional spaces.
- **Limitations:** Computationally intensive, sensitive to feature scaling.

**Model Parameters (Notebook Reference: Code Cell 8):**

- kernel='rbf'
- C=1000 (Penalty parameter of the error term)
- gamma=0.1 (Defines influence of a single training example)
- epsilon=0.2 (Defines epsilon-tube within which no penalty is associated)

### 🌲 What is Random Forest Regression (RFR)?

**Random Forest Regression** is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees.

- **Advantages:** Handles non-linearity, robust to outliers, low overfitting risk.
- **Limitations:** May require more memory and longer training time.

**Model Parameters (Notebook Reference: Code Cell 10):**

- n_estimators=100
- max_depth=None (nodes are expanded until all leaves are pure)
- random_state=42

---

### 4. Evaluation Results

The models were evaluated using the following metrics:

- **MAE (Mean Absolute Error)**
- **MSE (Mean Squared Error)**
- **RMSE (Root Mean Squared Error)**
- **$R^2$ Score (Coefficient of Determination)**

### 📊 Performance Comparison Table

Evaluation Metrics Results

| Metric | Random Forest | SVR |
|---|---|---|
| Mean Absolute Error (MAE) | 139.49 | 127.7 |
| Mean Squared Error (MSE) | 57561.69 | 54475.83 |
| Root Mean Squared Error (RMSE) | 239.92 | 233.4 |
| R-squared ($R^2$) | 0.182 | 0.226 |

---

### 5. Discussion & Conclusions

**Key Insights:**

- **Random Forest** outperformed SVR in all metrics, indicating better generalization and robustness to data noise.
- **SVR** required extensive tuning and preprocessing (especially scaling), and still did not match the performance of RF.
- Feature importance analysis from RF highlighted:
    - **RAM**, **Battery Capacity**, and **Internal Storage** as top contributors.
- Complex features like **Processor** and **Camera MP** were also influential, especially when combined with price tiers.

**Challenges:**

- The dataset lacked temporal or market-based features (e.g., release year, market trends).
- Some categorical features were overly simplified through encoding.

---

### 6. Final Remarks

This project successfully demonstrates the potential of machine learning in practical business use cases such as price prediction. By comparing SVR and Random Forest, we gain valuable insights into model behavior, preprocessing requirements, and practical trade-offs.