PROJECT IMPLEMENTATION

The implementation of this project follows a structured and rigorous data science workflow, transforming raw, noisy horological data into a high-precision predictive engine. The process is categorized into five major phases.

1. DATA LOADING AND EXPLORATORY ANALYSIS (EDA)

The journey began with the ingestion of the samiwatches.csv dataset. This phase was crucial for understanding the underlying data distribution:

- **STATISTICAL PROFILING:** Descriptive statistics were used to uncover price skews and the chronological age of brands. This revealed a significant variance in luxury pricing, necessitating robust preprocessing.
- **CORRELATION MAPPING:** A heatmap analysis was utilized to identify linear and non-linear relationships. This provided early evidence that while price is a strong indicator, technical specifications like water resistance and movement type also significantly influence prestige.

2. DATA REFINEMENT AND DOMAIN-BASED CLEANING

To ensure the model learns from high-quality signals, a meticulous cleaning and "feature surgery" phase was executed:

- **FEATURE PRUNING:** Columns identified as "Data Leakage" (variables created during initial data scraping that already contained information about the target, such as brand_avg_price) were removed. Redundant noise like URLs and detailed reference numbers were also pruned to prevent overfitting.
- **INTELLIGENT ENRICHMENT:** Missing brand_country data was handled through a dual-layered approach. First, an automated mapping was created from existing brand data. Second, manual domain-based imputation was used for independent horological houses (e.g., mapping *H. Moser & Cie.* to Switzerland), ensuring geographical prestige was correctly represented.
- **MULTILINGUAL STANDARDIZATION:** A custom mapping dictionary was developed to unify categorical features. Terms in Turkish (e.g., "Otomatik", "Çelik") and inconsistent English terms were merged into a standardized global taxonomy (e.g., "automatic", "stainless_steel").

3. ADVANCED PREPROCESSING AND ANOMALY DETECTION

Before modeling, the data underwent sophisticated treatment to eliminate biases:

- **ONE-HOT ENCODING:** Categorical variables were transformed into binary vectors. This was essential to prevent the model from assuming a mathematical order where none exists (e.g., treating "Black Dial" as numerically greater than "Blue Dial").
- **OUTLIER MANAGEMENT (ISOLATION FOREST):** To protect the model from "fraudulent" data points or scraping errors, an **Isolation Forest** was deployed. This algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

- **AUTOMATED CONTAMINATION TUNING (KNEED):** Unlike standard approaches that guess the outlier ratio, a **KneeLocator** was applied to the Isolation Forest decision scores. By identifying the "elbow" point of the score distribution, the "ideal contamination" threshold was mathematically determined, ensuring that only the most reliable training samples remained.

4. MODELING AND HYPERPARAMETER OPTIMIZATION

A **Multi-Model Pipeline Architecture** was established to rigorously benchmark different algorithmic approaches:

- **PIPELINE INTEGRATION:** Every model was wrapped in a scikit-learn Pipeline. For distance-based models (k-NN, SVM), a StandardScaler was included to ensure that large numerical values (like Price) did not overpower smaller ones (like Brand Age).
- **EXHAUSTIVE GRIDSEARCHCV:** A 5-fold cross-validation grid search was performed to fine-tune the "knobs" of each model:
    - **RANDOM FOREST:** Optimized for n_estimators (tree count) and max_depth to prevent the trees from simply memorizing the training data.
    - **SVM:** Tested various Kernels (RBF vs. Linear) and the C penalty parameter to find the optimal decision boundary.
    - **k-NN & NAIVE BAYES:** Refined neighbor weights and smoothing constants to handle class boundaries.
- **SCORING STRATEGY:** The models were optimized using the **F1-Weighted Score**. This metric was chosen specifically because it accounts for class imbalance, ensuring the model is as good at identifying rare "Ultra-Luxury" watches as it is at identifying common "Entry-Level" ones.

5. EVALUATION AND PERFORMANCE BENCHMARKING

Final validation was conducted on a "hold-out" test set that the model had never seen before:

- **CONFUSION MATRIX ANALYSIS:** This allowed for a granular view of misclassifications, ensuring that the model does not make "catastrophic" errors (e.g., confusing an Entry-Level watch with an Ultra-Luxury one).
- **ROC-AUC CURVES:** Multi-class ROC curves were plotted to measure the sensitivity and specificity. The high AUC values across all classes proved the models' exceptional "separability" power.
- **WINNER IDENTIFICATION: Random Forest** was crowned the champion. It achieved a near-perfect **98.8% F1-Weighted Score** on both training and test sets. This remarkable consistency indicates that the model has successfully "decoded" the complex logic of watch prestige, making it a robust tool for real-world horological analysis.