Project Report

I. Introduction

Project Goal:

The primary objective of this capstone project is to develop and evaluate a comprehensive machine learning pipeline addressing two distinct real-world problems. First, we aim to build a **Regression Model** to predict continuous housing prices using the California Housing dataset. Second, we aim to construct a **Classification Model** to assist in the early diagnosis of heart disease by predicting patient risk levels. Additionally, Unsupervised Learning (Clustering) is applied to uncover hidden patterns within the data.

Data Description:

- **Source:** We utilized two publicly available datasets sourced from **Kaggle**: the **Heart Disease Prediction Dataset** for the classification task and the **California Housing Dataset** for the regression task.
- **Size:**
  - The **Heart Disease Dataset** consists of **2,181 observations (rows)** and **14 columns**.
  - The **California Housing Dataset** contains census block groups with continuous numerical records suitable for regression analysis.
- **Features:**
  - **Heart Disease:** Includes **13 independent clinical features** such as **Age**, **Sex**, **Chest Pain Type (cp)**, **Resting Blood Pressure (trestbps)**, and **Cholesterol (chol)**. The target variable is binary (**0**: Healthy, **1**: Disease).
  - **California Housing:** Comprises continuous features representing housing metrics (e.g., **Median Income**, **Total Rooms**), with **Median House Value** serving as the target variable for prediction.

II. Data Preprocessing & EDA

Data Cleaning Steps:

- **Missing Values:**

    - **Heart Disease Dataset:** Addressed data quality issues where missing values were represented by **'?'** characters. These were converted to NaN via coercion (pd.to_numeric). To maintain data integrity, we imputed these missing values using the **Median** strategy, which is robust against outliers in physiological features like cholesterol.
    - **Housing Dataset:** Identified **207 missing values** in the total_bedrooms column. These were imputed using the **Mean** value of the column, as the distribution allowed for average-based filling without significant distortion.
- **Encoding:**
    - **Heart Disease Dataset:** Categorical features such as sex, cp, and thal were already present in a numerical label-encoded format. Therefore, no additional One-Hot Encoding was required for the models utilized.
    - **Housing Dataset:** All features were continuous numerical variables (floats/ints), requiring no categorical encoding.
- **Other Transformations:**
    - **Feature Scaling:** Applied **StandardScaler** to both datasets. This normalized the data (Mean=0, SD=1), ensuring that features with large ranges (e.g., chol > 200 or total_rooms > 1000) did not bias distance-based algorithms like KNN and K-Means.
    - **Outlier Detection (Housing):** Visualized the median_income feature using boxplots. While outliers were detected, they were interpreted as valid high-income data points rather than errors, and thus were retained to preserve real-world variance.
    - **Type Conversion (Heart):** Forced conversion of object-type columns (e.g., trestbps, chol) to float64 to resolve the issue of non-numeric characters preventing analysis.

Key Findings from EDA:

- **Correlations:**Exploratory analysis suggests a strong positive correlation between **median_income** and housing prices. The boxplot for median income revealed distinct outliers in the upper brackets, indicating that higher-income demographics are a primary driver for variance in house values.
- **Target Distribution:**
    - The target variable is a **binary classification** label (**0**: Healthy, **1**: Disease). The analysis confirmed that the dataset contains distinct labeled examples for both classes, validating the suitability of using binary classification algorithms (like Logistic Regression) rather than

III. Modeling and Results

Regression Results:

- **Model Used:** Linear Regression was used to model the relationship between the input features and the target variable (median_house_value). This model was chosen as a baseline regression algorithm due to its simplicity and interpretability.
- **Final MSE/RMSE:**
    - The Linear Regression model achieved a Mean Squared Error (MSE) of 5,052,953,703 and a Root Mean Squared Error (RMSE) of 71,084.
    - The RMSE value indicates that, on average, the model's predicted house prices deviate from the actual house prices by approximately 71,000 units, which provides a clear and

interpretable measure of the prediction error in the same unit as the target variable.

- **R² Interpretation:**
  - The model achieved an R² score of 0.61, meaning that approximately 61% of the variance in house prices is explained by the input features used in the regression model.
  - This result suggests that the Linear Regression model is able to capture a substantial portion of the relationship between the features and the target variable, while the remaining unexplained variance may be due to factors not included in the dataset or the linear nature of the model.

Classification Results:

- **Model Used:**
  - Two classification models were implemented and evaluated in this project:
    - Logistic Regression
    - Decision Tree Classifier
  - Both models were trained on the scaled training dataset and tested on unseen data to classify whether a patient has heart disease (1) or not (0).
- **Final Accuracy:**
  - Logistic Regression Accuracy: 75.29% .This means the model correctly classified approximately 75 out of every 100 patients.
  - Decision Tree Accuracy: 94.5%. This means the Decision Tree model correctly classified approximately 95 out of every 100 patients, making it the better-performing model in terms of overall accuracy.
- **Confusion Matrix Analysis:**
  - Logistic Regression Confusion Matrix
    - True Negatives (151): Patients without heart disease correctly classified.
    - False Positives (66): Healthy patients incorrectly classified as having heart disease.
    - False Negatives (42): Patients with heart disease incorrectly classified as healthy.
    - True Positives (178): Patients with heart disease correctly classified.
    - **Interpretation**: False negatives are particularly critical in this medical context, as they represent patients who have heart disease but were not detected by the model. Logistic Regression produced 42 false negatives, which indicates a moderate risk of missed diagnoses.
  - Decision Tree Confusion Matrix
    - True Negatives (201): Healthy patients correctly classified.
    - False Positives (16): Healthy patients incorrectly classified as diseased.
    - False Negatives (8): Patients with heart disease incorrectly classified as healthy.
    - True Positives (212): Patients with heart disease correctly classified.
    - **Interpretation**: The Decision Tree model reduced false positives compared to Logistic Regression, resulting in fewer unnecessary alarms for healthy patients.
  - Overall Discussion
    - The Decision Tree model achieved higher overall accuracy and fewer false positives.
    - The final model choice depends on whether minimizing missed diagnoses (false negatives) or maximizing overall accuracy is the primary objective.

Clustering Results:

- **Chosen K:** The Elbow Method shows a clear bend at K = 3, after which the decrease in inertia becomes marginal. This choice is further supported by the Silhouette Analysis, where K = 3 achieves the highest silhouette score. Therefore, K = 3 was selected as the optimal number of clusters.
- **Silhouette Score:** The clustering achieved a silhouette score of 0.186, indicating a good level of separation between clusters and meaningful group structure in the data.
- **Cluster Interpretation:**
    - The clustering results reveal three distinct patient groups with varying cardiovascular risk profiles. Despite moderate overlap, the clusters highlight meaningful relationships between age and maximum heart rate, underscoring the value of unsupervised learning for exploring patterns in medical datasets..

IV. Conclusion

Best-Performing Model Summary for each Task:

- Regression: Linear Regression – good RMSE and R^2, effective for predicting house prices.
- Classification: Decision Tree – 94.5% accuracy, fewer false positives.
- Clustering: K-Means (K=3) – clear, homogeneous clusters for pattern analysis.

Challenges and Future Work:

- Challenges: missing/inconsistent data, choosing metrics, feature scaling variability.
- Future Work: try advanced models (Gradient Boosting, Random Forest), use PCA, handle class imbalance, improve interpretability.

Part 4: Reporting and Submission

9. Final Report Generation

Github : GitHub - omar3laa/ML-Capstone-Project