**Project AI**

**Team members :  
  
Khaled Abuelenein 20224609  
Gamal Mohammed 20223482  
Seraj Mokhtar 20224142  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Documentation: Heart Disease Prediction Project**

**Project Goals**

The main objective of this project is to predict heart disease presence using machine learning models. The project involves extensive data preprocessing, feature selection, handling imbalanced datasets, and applying supervised and unsupervised learning techniques. Additionally, the project's outcome aims to evaluate the performance of models and provide visualizations to explain results.

**Steps, Methods, and Challenges Addressed**

**1. Data Understanding and Exploration**

* **Dataset Used**: heart\_disease\_uci.csv
* **Initial Dataset Shape**: (rows, columns)
* **Missing Data**:
  + Total missing values: Checked and imputed where necessary.
  + Methods used:
    - Categorical columns: Imputed using mode.
    - Numeric columns: Imputed using mean.
* **Duplicates**:
  + Total duplicates: Found and removed to clean the data.
* **Outliers**:
  + Z-score method with a threshold of 3.
  + Rows with outliers removed.

**2. Data Cleaning and Preprocessing**

* **Categorical Encoding**: Applied LabelEncoder to transform categorical variables.
* **Scaling**: Used MinMaxScaler to normalize numeric columns.
* **Principal Component Analysis (PCA)**:
  + Reduced dimensionality to 2 principal components.
  + Compared model performance with and without PCA.
* **Imbalanced Dataset Handling**:
  + Applied SMOTE and RandomOverSampler.
  + Verified class distribution before and after resampling.

**3. Feature Selection**

* **Selected Features**: ['age', 'chol', 'thalch', 'oldpeak', 'ca']
* **Correlation Heatmap**: Visualized correlations for selected features and overall features to confirm minimal multicollinearity.

**4. Machine Learning Models**

**Supervised Learning**

1. **Logistic Regression**
   * Compared performance with all features, selected features, PCA, and original dataset.
2. **Random Forest**
   * Evaluated with selected features and SMOTE.
3. **Support Vector Machines (SVM)**
   * Assessed performance with and without resampling.

**Unsupervised Learning**

* **K-Means Clustering**
  + Used the Elbow method to determine the optimal number of clusters.

**Challenges and Solutions**

1. **Imbalanced Dataset**:
   * Challenge: Unequal distribution of target classes.
   * Solution: Resampling with SMOTE and RandomOverSampler.
2. **Outliers**:
   * Challenge: Influenced mean and skewed the dataset.
   * Solution: Used Z-score to remove significant outliers.
3. **Feature Selection**:
   * Challenge: High dimensionality caused noise in the dataset.
   * Solution: PCA and correlation analysis to reduce dimensionality.

**Summary Table: Data Cleaning and Preprocessing Techniques**

|  |  |  |
| --- | --- | --- |
| Technique | Applied (True/False) | Handling Method |
| Missing Data Imputation | True | Mode for categorical, Mean for numeric |
| Duplicates Removal | True | Removed duplicates |
| Outliers Detection | True | Z-score method |
| Scaling | True | MinMaxScaler |
| Dimensionality Reduction | True | PCA |
| Handling Imbalanced Dataset | True | SMOTE, RandomOverSampler |

**Project Goals**

The primary goal of this project is to predict the likelihood of heart disease in patients using machine learning techniques. The dataset used for this analysis includes various patient features such as age, cholesterol levels, blood pressure, and other clinical measures. The project employs a range of techniques including data preprocessing, dimensionality reduction, feature selection, resampling, supervised and unsupervised learning, and model evaluation. The ultimate aim is to build accurate models for predicting heart disease and understand the factors contributing to the condition.

Methodology and Techniques Used

**1. Data Preprocessing**

Handling Missing Values: The dataset contains missing values, which are handled differently based on the type of column:

Categorical columns (e.g., fbs, restecg, exang, slope, thal, ca) are imputed using the mode (most frequent value).

Numeric columns (e.g., trestbps, chol, thalch, oldpeak) are imputed using the mean of the respective columns.

Removing Duplicates: Duplicate entries are checked using duplicated() and removed to avoid redundancy in the data.

Handling Outliers: Outliers in the numeric columns are detected using Z-scores, and rows containing outliers are removed if their Z-score exceeds a threshold of 3.

**2. Feature Encoding and Scaling**

Label Encoding: Categorical features such as gender, dataset, cp, fbs, restecg, exang, slope, and thal are transformed into numerical values using LabelEncoder. This ensures that categorical data can be used in machine learning models.

Feature Scaling: Numeric columns like age, trestbps, chol, thalch, oldpeak, ca, and num are scaled using MinMaxScaler to ensure that all features are within the range [0, 1], which helps improve model convergence.

**3. Dimensionality Reduction (PCA)**

Principal Component Analysis (PCA): PCA is applied to reduce the dimensionality of the dataset while preserving as much variance as possible. In this project, PCA is applied with 2 components to reduce the feature space for visualization and model training. This technique helps in identifying the most important features and improving model performance.

**4. Handling Class Imbalance**

SMOTE (Synthetic Minority Over-sampling Technique): SMOTE is used to generate synthetic samples for the minority class, thus balancing the dataset. It improves model performance by addressing the issue of class imbalance.

Random OverSampling: An alternative technique to SMOTE, RandomOverSampler is applied to balance the classes by randomly duplicating examples from the minority class.

**5. Modeling Techniques**

Several machine learning models are applied to predict heart disease, including:

Supervised Learning:

Support Vector Machine (SVM): A classification model that finds the optimal hyperplane separating different classes in high-dimensional spaces.

Random Forest Classifier: An ensemble model that combines multiple decision trees to improve accuracy and prevent overfitting.

Logistic Regression: A linear model used for binary classification that estimates the probability of the target class.

Unsupervised Learning:

K-Means Clustering: This algorithm is used for grouping similar instances into clusters. PCA-transformed data is used to apply K-Means and visualize clusters.

**6. Model Evaluation**

The models are evaluated using a variety of metrics to assess their performance:

Accuracy: The percentage of correct predictions.

Precision: The proportion of true positive predictions out of all positive predictions.

Recall: The proportion of true positive predictions out of all actual positive cases.

F1-Score: The harmonic mean of precision and recall, used when dealing with imbalanced datasets.

Confusion Matrix: A matrix showing the true positives, false positives, true negatives, and false negatives for model evaluation.

Log Loss: A measure of the uncertainty of the model’s predictions.

The performance is evaluated using these metrics both with and without resampling techniques like SMOTE and RandomOverSampler.

**7. PCA-Transformed Data Evaluation**

Model Comparison: The models are evaluated both with and without PCA transformation to compare their performance. This helps determine whether reducing dimensionality improves the model’s ability to predict heart disease.

**8. Correlation Analysis**

Correlation Heatmaps: Correlation matrices are generated to understand the relationships between different features. This analysis helps identify which features are highly correlated and can be used to select important variables for model training.

Selected Features: Features such as age, chol, thalch, oldpeak, and ca are selected based on correlation and domain knowledge for model training. This is intended to reduce dimensionality and improve model efficiency.

**9. Clustering and Further Analysis**

Clustering with K-Means: K-Means clustering is used on the PCA-transformed dataset to identify natural groupings in the data. This unsupervised learning technique provides insights into the structure of the dataset, which might reveal patterns not captured by the classification models.

**Model Evaluation and Results**

**1. Logistic Regression**

* **With All Features**:
  + Accuracy: 47%
* **With Selected Features**:
  + Accuracy: 76%

**2. Random Forest With SMOTE Resampling**:

* + Accuracy: 90%

**3. SVM Without Resampling**:  
 accuracy=60%

**4. Metrics for Logistic Regression WITH SMOTE** Accuracy: **48.19%**

* + Precision: **46.07%**
  + Recall: **48.19%**
  + F1-Score: **46.40%**
  + Log Loss: **1.1984**

**5. Metrics for SVM WITH RandomOverSampler**

* + Accuracy: **60.96%**
  + Precision: **61.04%**
  + Recall: **60.96%**
  + F1-Score: **60.73%**

**6. Metrics for Random Forest WITH RandomOverSampler**

* + Accuracy: **90.36%**
  + Precision: **90.33%**
  + Recall: **90.36%**
  + F1-Score: **90.18%**

**Unsupervised Learning**

* **K-Means Clustering**:
  + Elbow Method applied.
  + Optimal clusters determined: X clusters.

**Visualizations**

1. **Correlation Heatmap**
   * Visualized correlation of numeric features.
   * Used heatmaps before and after feature selection.
2. **Elbow Method Chart for K-Means Clustering** Elbow Chart Placeholder]
3. **Model Performance Comparisons**
   * Charts comparing accuracy, precision, recall, and F1-score for models under different preprocessing techniques.

![A graph of blue and orange bars

Description automatically generated