ALBUKHARY INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTING & INFORMATICS

Course Title : MACHINE LEARNING
Course Code : CCS2213
Lecturer : Prof Dr Zurinahni binti Zainol
Semester/Year : 2, 2023/2024
Submission Date : 27 June 2024
Assignment Weightage : 20%

PROJECT Assessment

CLO4: Develop machine learning models (C6, PLO6)

Name	Student ID
Thinley Yeshey Choden	AIU22102188

Comments

Lecturer's Name: Prof Dr Zurinahni binti Zainol

Table of content

Introduction **3**
Objectives and Background **3**
Methodology1 ..... **4**
1. Decision Tree Classifier **5**
2. Support Vector Machine (SVM) **6**
3. K-Nearest Neighbors (KNN) **6**
4. Logistic Regression **7**
5. Unsupervised Learning with K-Means Clustering **7**
Methodology2 ..... **8**
1. K-means Clustering 9
2. Hierarchical Clustering 10
Results and Discussion **10**
Conclusion and Future works **15**
References **16**

Introduction

This project integrates machine learning techniques to address critical challenges in healthcare and global development. It focuses on predicting heart failure outcomes using clinical data and clustering countries based on socio-economic indicators. By evaluating supervised learning models for healthcare predictions and applying unsupervised learning for country clustering, the project aims to enhance treatment strategies and optimize resource allocation in humanitarian efforts.

Objectives and Background

The report consists of two datasets focused on different objectives.

1. **Healthcare Predictive Analytics:**

In the realm of healthcare predictive analytics, the project employs a suite of supervised learning algorithms tailored to predict heart failure outcomes based on clinical data:

Supervised Learning Algorithms:

- Decision Trees: Utilized to create predictive models that segment patient data based on hierarchical decisions, providing insights into risk factors such as age, ejection fraction, and serum creatinine levels.
- Support Vector Machines (SVM): Applied for classifying patients into survival groups based on complex decision boundaries derived from clinical features.
- K-Nearest Neighbors (KNN): Used to predict patient outcomes by comparing their features to those of similar patients in the dataset.
- Logistic Regression: Employed to model the probability of heart failure occurrence based on explanatory variables like age, sex, and medical history.

Unsupervised Learning Algorithms:

- K means algorithm: cluster based on binary label 0 and 1 for death.

2. Humanitarian Aid Allocation:

The project also addresses humanitarian aid allocation through the application of unsupervised learning algorithms on socio-economic data from diverse countries:

Unsupervised Learning Algorithms:

- K-Means Clustering: Utilized to categorize countries into clusters based on socio-economic indicators such as GDP, health expenditure, and education levels, aiming to identify groups with similar developmental needs.
- Agglomerative Hierarchical Clustering: Provides a hierarchical decomposition of the dataset into clusters, revealing nested structures and facilitating nuanced insights into global development disparities.

Methodology1

Description of the Dataset:

The dataset used for this study is the Heart Failure Clinical Records Dataset, sourced from the Faisalabad Institute of Cardiology and the Allied Hospital in Pakistan. It comprises 299 instances and 13 features encompassing physical, clinical, and lifestyle information of patients with previous heart failures. Key features include age, ejection fraction, serum creatinine, and the binary target variable, DEATH_EVENT, indicating patient survival status during the follow-up period.

1. Data Preprocessing

Loading the Data:

Utilizing pandas, the dataset was loaded and initially inspected using df.head(), df.info(), and df.describe() to comprehend its structure and content.

2. Exploratory Data Analysis (EDA):

Histograms and Boxplots: Visualizations were employed to explore distributions and relationships between features. For instance, examining age distribution across survival outcomes and plotting ejection fraction vs. serum creatinine to discern patterns related to patient outcomes.

Correlation Analysis: A heatmap of the correlation matrix was generated to identify significant correlations between features and the DEATH_EVENT target variable. Notably, variables such as serum creatinine and age showed positive correlations with mortality, while features like ejection fraction exhibited negative correlations.

Handling Missing Values:
The dataset was verified for missing values using df.isna().sum(), revealing a well-constructed dataset with no null values, thus requiring no further cleaning steps.

3. Feature Scaling:

Binary and Non-binary Data: Features were categorized into discrete and continuous types based on their nature. While discrete features (e.g., diabetes, high blood pressure) required no scaling due to their binary nature, continuous features (e.g., age, serum creatinine) were scaled using MinMaxScaler to normalize their ranges, ensuring no feature disproportionately influenced model training.

4. Data Splitting:
The dataset was split into training (70%) and testing (30%) sets using train_test_split, stratifying by DEATH_EVENT to maintain class distribution integrity.

Supervised Learning Algorithms

1. Decision Tree Classifier

1. **Model Training**:

- A Decision Tree Classifier was trained on the training data. The model's maximum depth was varied from 1 to 8 to understand the impact of tree depth on model performance.

2. **Evaluation**:
    - The model was evaluated using training accuracy, testing accuracy, and 5-fold cross-validation accuracy for each depth level.
    - The results were recorded and plotted to visualize the performance.
3. **Visualization**:
    - A decision tree with a depth of 3 was visualized to understand the decision rules and structure of the tree.

2. Support Vector Machine (SVM)

1. **Hyperparameter Tuning**:
    - Hyperparameter tuning was conducted using GridSearchCV and RandomizedSearchCV to find the best parameters for the SVM. The parameters tuned included kernel type, C (regularization parameter), degree (for polynomial kernel), and gamma (kernel coefficient).
2. **Model Training and Evaluation**:
    - The best parameters were identified, and the model was trained using these parameters.
    - The model was evaluated using cross-validation accuracy and test accuracy.
    - A confusion matrix was plotted to visualize the model's performance on the test set.

3. K-Nearest Neighbors (KNN)

1. **Hyperparameter Tuning**:
    - A GridSearchCV was used to find the optimal hyperparameters for the KNN classifier. The parameters tuned included the number of neighbors (n_neighbors), weights (uniform or distance), and distance metric (euclidean, manhattan, minkowski).

2. **Model Training and Evaluation**:
    - The best parameters were identified, and the model was trained using these parameters.
    - The model was evaluated using cross-validation accuracy and test accuracy.
    - A confusion matrix and classification report were generated to assess the model's performance.
3. **Visualization**:
    - The cross-validation accuracy for different values of n_neighbors was plotted to visualize how the number of neighbors affects the model performance.

4. Logistic Regression

1. **Hyperparameter Tuning**:
    - A GridSearchCV was used to find the optimal hyperparameters for the Logistic Regression model. The parameters tuned included the regularization parameter (C) and penalty type (L2).
2. **Model Training and Evaluation**:
    - The best parameters were identified, and the model was trained using these parameters.
    - The model was evaluated using cross-validation accuracy and test accuracy.
    - A confusion matrix and classification report were generated to assess the model's performance.

5. Unsupervised Learning with K-Means Clustering

1. **Clustering**:
    - The K-Means algorithm was used to cluster the data into 2 clusters.
    - The elbow method was employed to determine the optimal number of clusters by plotting the loss (inertia) for a range of cluster numbers.
2. **Evaluation**:
    - The clustering results were evaluated by comparing the cluster labels with the actual `DEATH_EVENT` labels using a contingency table and calculating accuracy.

3. **Visualization**:
    - The clustered data points and cluster centroids were plotted to visualize the K-Means clustering results.

Methodology2

Description of the dataset:

Unsupervised Learning on Country Data:

Fund Allocation for Countries in Need HELP International is an international humanitarian NGO dedicated to combating poverty and providing essential relief during disasters and natural calamities. Having raised around $10 million, the organization needs to allocate these funds strategically and effectively. To determine which countries are in the direst need of aid, data-driven decisions must be made. By categorizing countries using socio-economic and health factors that determine overall development, clusters can be formed to guide fund allocation. This is a classic case of unsupervised learning, where clusters are created based on various features.

Aim

The aim is to cluster countries based on numerical features related to socio-economic and health indicators, enabling effective fund allocation. This unsupervised learning problem statement involves using K-means and hierarchical clustering techniques.

Dataset Attributes

- country: Name of the country
- child_mort: Deaths of children under 5 years of age per 1000 live births
- exports: Exports of goods and services per capita as a percentage of GDP per capita
- health: Total health spending per capita as a percentage of GDP per capita
- imports: Imports of goods and services per capita as a percentage of GDP per capita
- Income: Net income per person
- Inflation: Annual growth rate of the total GDP
- life_expec: Average number of years a newborn child would live if current mortality patterns continue

- total_fer: Number of children that would be born to each woman if current age-specific fertility rates remain the same
- gdpp: GDP per capita, calculated as the total GDP divided by the total population

1. Data Preparation and Exploration
1. Loading and Inspecting Data: The dataset, "Country-data.csv," was loaded using pandas, and its structure was inspected to ensure data integrity and readiness for analysis.
2. Visualizing Distributions: Distribution plots for each numerical feature were generated to identify skewness and outliers, providing a clear understanding of the data distribution.
3. Exploring Top and Bottom Countries by Key Indicators: Bar plots highlighted countries with extreme values for key indicators like child mortality, exports, health spending, and GDP per capita, offering insights into the data extremes.
4. Correlation Analysis: A correlation matrix was plotted to identify relationships between features, aiding in effective feature engineering and clustering.
5. Feature Engineering and Normalization

Based on the EDA, three composite features were engineered:

- Health: Combined child mortality, health spending, life expectancy, and fertility rates.
- Trade: Included imports and exports.
- Finance: Encompassed income, inflation, and GDP per capita.

6. These features were normalized using MinMaxScaler to ensure they are on a comparable scale, facilitating effective clustering.
7. Clustering Techniques

1. K-means Clustering

To determine the optimal number of clusters for K-means clustering, the elbow method and silhouette scores were used:

- **Elbow Method:** The sum of squared errors (SSE) was plotted for a range of cluster numbers to identify the optimal number of clusters where the SSE starts to level off.
- **Silhouette Scores:** Evaluated how similar each point is to its own cluster compared to other clusters, with higher scores indicating better-defined clusters.

Three clusters were chosen as the optimal number. The clustering results were visualized in 3D and 2D plots, highlighting the distinct characteristics of each cluster.

2. Hierarchical Clustering

Hierarchical clustering was performed using the Ward linkage method, minimizing variance within each cluster. The optimal number of clusters was determined using:

- **Dendrogram:** Displayed the arrangement of clusters formed at each step, helping identify the number of clusters.
- **Silhouette Scores:** Evaluated the quality of the clusters.

Three clusters were chosen based on the dendrogram and silhouette scores. The results were visualized in 3D and 2D plots to provide insights into the cluster characteristics.

Comparison of Clustering Results

Silhouette scores for both K-means and hierarchical clustering were compared to evaluate which method produced more distinct and well-defined clusters.

Results and Discussion

1. **Supervised** on heart failure clinical records:

Table 1: Results on Heart failure Dataset

Models	Accuracy
Decision Tree	0.8470
SVM	0.8280
KNN	0.7515
Logistic regression	0.8282
K-means	0.5552

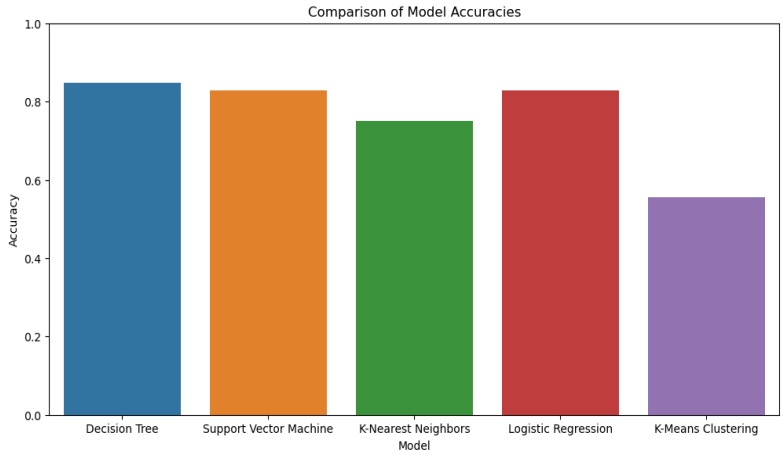

Figure 1: Comparative analysis on supervised and unsupervised models on Heart Failure Dataset.

The result of the supervised algorithm modeling after fine tuning and optimization proved Decision Tree to be the best model on the dataset with the highest accuracy of approximately 0.84%.

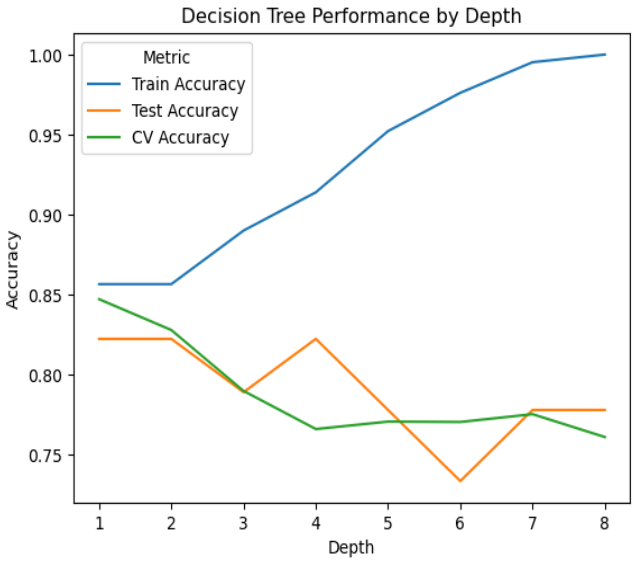

Figure 2: Decision tree Performance

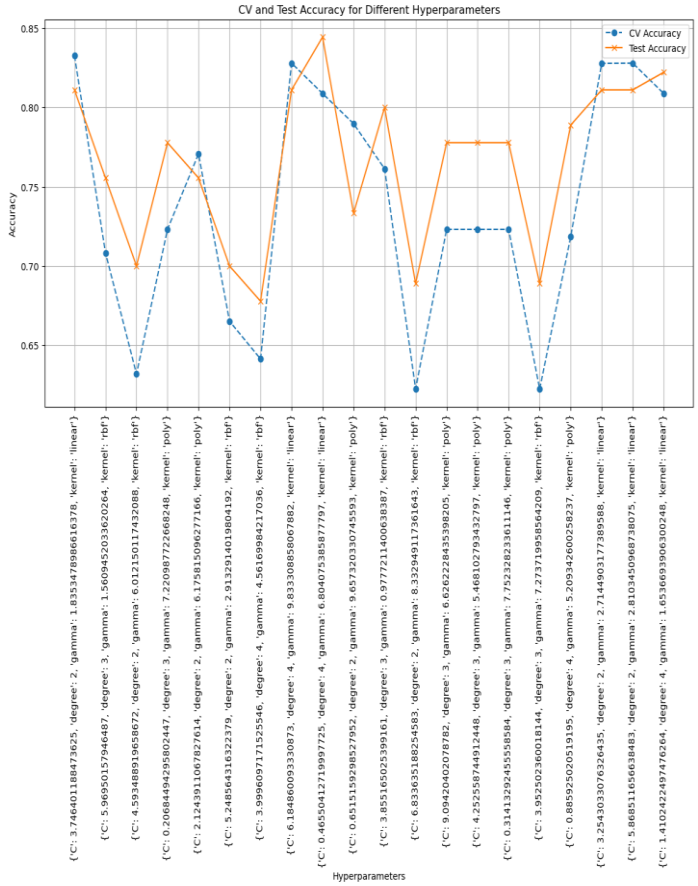

Figure 3: Accuracy for SVM

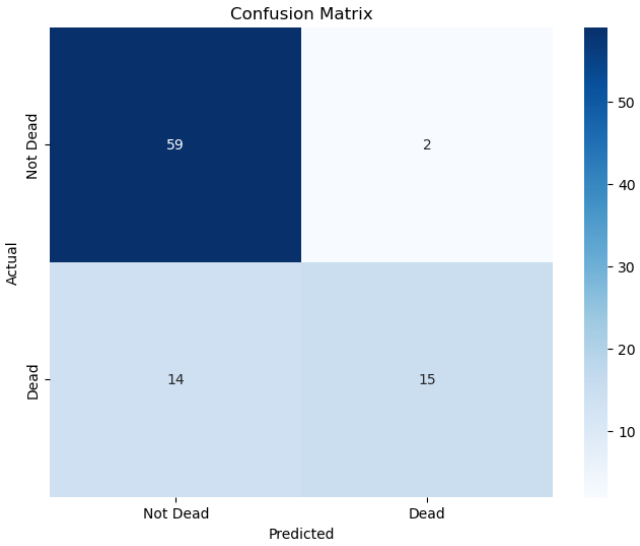

Figure 4: Confusion matrix for KNN

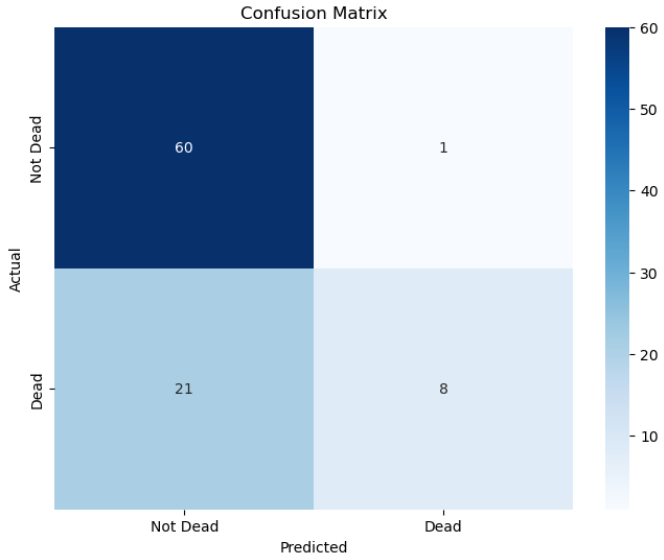

Figure 5: Confusion matrix for Logistic Regression

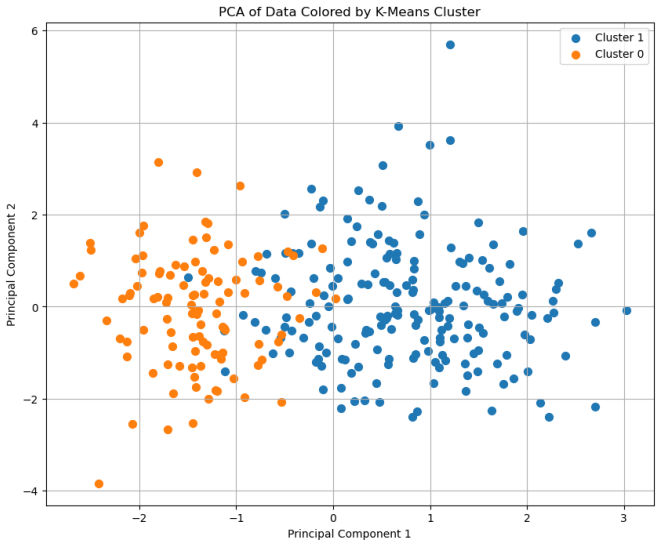

Figure 6: k means clustering

2. **Unsupervised** on heart failure clinical records:

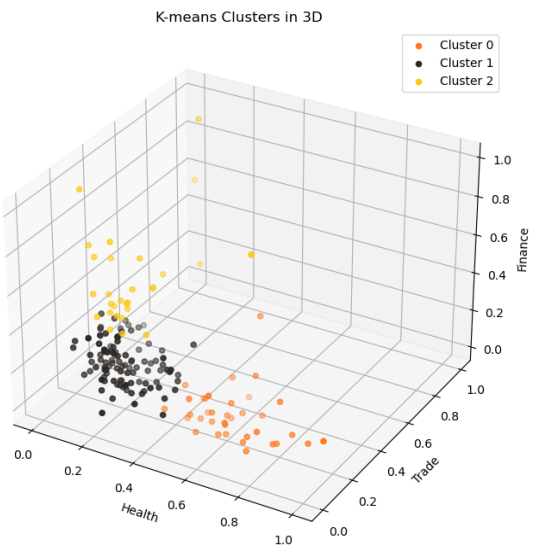

Figure 7: clustering for k means

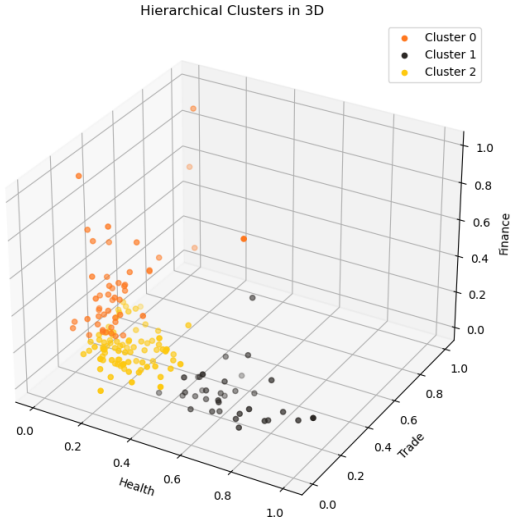

Figure 8: clustering for hierarchical

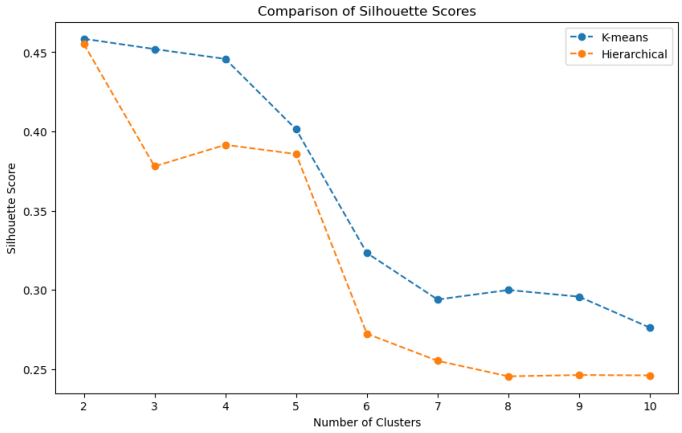

Figure 9: comparison of Silhouette score

Table 2: Silhouette scores.

Number of clusters	K means Silhouette Score	Hierarchical Silhouette Score
2	~0.45	~0.45
3	~0.45	~0.375

4	~0.45	~0.4
5	~0.4	~0.39

The table shows silhouette scores for K-Means and Hierarchical Clustering across different numbers of clusters.
Scores indicate that 2 or 3 clusters are optimal, with clarity decreasing as the number of clusters increases.
K-Means generally outperforms Hierarchical Clustering in maintaining higher silhouette scores, suggesting clearer cluster definitions in this dataset.

Conclusion and Future works

In summary, this project has demonstrated the efficacy of machine learning in healthcare and humanitarian aid allocation. Supervised learning models, particularly the Decision Tree, showed promising results in predicting heart failure outcomes with an accuracy of approximately 84%. Unsupervised learning techniques, specifically K-Means clustering, effectively categorized countries based on socio-economic indicators, offering insights crucial for targeted resource allocation.

The implications of these findings are significant for both medical prognosis and strategic humanitarian initiatives, highlighting the potential to optimize healthcare interventions and enhance global development efforts through data-driven decision-making.

Future work could focus on refining model performance through more extensive feature engineering, exploring ensemble methods for enhanced predictive power, and integrating real-time data streams for dynamic decision support systems in healthcare and humanitarian logistics. These advancements promise to further leverage machine learning's potential in addressing complex societal challenges.

References

https://www.kaggle.com/code/nichyhan/heart-failure-dt-rf-svm-pca-k-means

https://www.kaggle.com/code/chandrimad31/clustering-with-pca-kmeans-hierarchical-dbscan