**MLUL2 - TECHNICAL REPORT**

**Group 1**

|  |  |
| --- | --- |
| **Group Member** | **PGID** |
| Avinav Anand | 12410041 |
| Jerry Mavely | 12410044 |
| Preksha Rai | 12410063 |
| Abhhay Verma | 12410081 |

# **Introduction:**

Heart disease remains one of the leading causes of death worldwide, making it crucial to explore factors that contribute to its risk. This project focuses on analysing the UCI Heart Disease dataset using unsupervised machine learning techniques such as anomaly detection algorithms, association rules etc. to uncover patterns, relationships, and anomalies in patient data.

The goal is to use methods like clustering, Principal Component Analysis (PCA), anomaly detection, and association rules to better understand the data and identify risk factors. Key health parameters such as cholesterol levels, blood pressure, heart rate, chest pain, and ECG results are analysed to gain meaningful insights.

The project's goals, procedures, and outcomes are described in this study, which also shows how unsupervised learning can be used to interpret complicated medical data and possibly support the risk assessment for heart disease.

## Problem Statement:

**Project statement:** Heart disease data exploration to draw conclusion based on patterns and anomalies in the dataset.

**Github Link:** <https://github.com/prai2025S/MLUL2_Group01.git>

# **Data and data cleaning:**

**Dataset Source:** UCI Heart Disease

**Link:**[Heart Disease Data Set from UCI data repository](https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data)

**EDA Code File Name:** *MLUL2-EDA.ipynb*

**Description of the dataset is given below:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Attribute** | **Type of data** | **Definition** |
| 1 | id | numeric | Unique id for each patient |
| 2 | age | numeric | Age of the patient in years |
| 3 | sex | categorical | Male/Female |
| 4 | dataset | categorical | place of study |
| 5 | cp chest pain type | categorical | typical angina, atypical angina, non-anginal, asymptomatic |
| 6 | trestbps | numeric | resting blood pressure (in mm Hg on admission to the hospital |
| 7 | chol | numeric | serum cholesterol in mg/dl |
| 8 | fbs | categorical | True/False based on if fasting blood sugar > 120 mg/dl |
| 9 | restecg | categorical | resting electrocardiographic results) -- Values: [normal, stt abnormality, lv hypertrophy |
| 10 | thalach | numeric | maximum heart rate achieved |
| 11 | exang | categorical | exercise-induced angina (True/ False) |
| 12 | oldpeak | numeric | ST depression induced by exercise relative to rest |
| 13 | slope | categorical | the slope of the peak exercise ST segment |
| 14 | ca | ordinal | number of major vessels (0-3) coloured by fluoroscopy |
| 15 | thal | categorical | Type of heart disorder: [normal; fixed defect; reversible defect] |
| 16 | num | ordinal | the severity of the heart disease on the scale of 0-4 |

## Data Cleaning:

**Data Cleaning Process:**

1. Data exploration: In first step of data cleaning, we explored the dataset and checked the data structure and data type of the values of each column.
2. Then we checked for duplicate values in data set. There were no duplicate values in this data set.
3. Handling Missing Values: We checked for missing values and calculated the null percentage for each column. The columns such as slope, ca and thal had null percentage as 33.58%, 66.41%, 52.82% respectively.
4. Imputation of Missing Values:
   * + Categorical Fields (*fbs*, *restecg*, *exang*): Imputed using the mode.
     + Numerical Fields:
     + Columns with outliers (*trestbps*, *chol*, *oldpeak*): Imputed with the median.
     + Columns without outliers (*thalch*): Imputed with the mean.
5. On further analysis, we decided that imputation with mean, median and mode can lead to bias in dataset and may affect out further analysis with anomaly detection and Association rules, so we decided to do KNN imputation.
6. KNN Imputation: Applied to handle remaining missing data by considering relationships between variables for accurate imputations:
   * + - The categorical and numerical columns were separated and treated accordingly.
       - Categorical columns were one hot encoded
       - For the numerical columns, first the 0’s in ‘chol’ and ‘trestbps’ were replaced by NaN. This was done to impute these values later. And then the numerical columns were scaled (using MinMaxScaler).
       - KNN imputation method was used to fill in the missing and Null values in the dataset and all values were saved as integers.

# **Empirical Analysis:**

## **Anomaly Detection:**

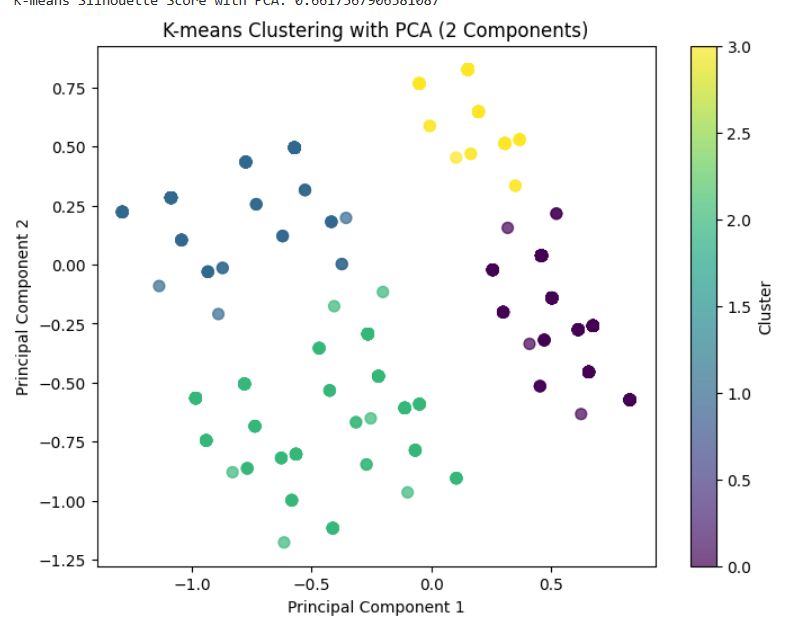
**Clustering Code File Name:** MLUL2 - Clustering & Anomaly Detection using LOF.ipynb

Anomalies were analyzed using Isolation Forest and Local Outlier Factor, focusing on features like cholesterol, resting blood pressure, heart rate, and ST depression. Both methods were applied on clusters generated through the following steps:

1. Initial Clustering:
   * Goal: Analyse patterns using K-Means and Hierarchical Clustering.
   * Evaluation: Low average silhouette scores from the first clustering results showed inadequate separation because of redundant features.
2. Problem Identification:
   * Issues: Clusters were poorly defined because of redundant features
   * Solution: Preprocessing was required to lower dimensionality and noise.
3. Feature Selection:
   * Method: Identified top 7 features by using Permutation Importance, these features contributed most to cluster formation.
   * Outcome: Reduced dimensionality eliminated noise and retained meaningful features.
4. Dimensionality Reduction Using PCA:
   * Goal: Simplify data and improve clustering.
   * Method: PCA reduced the dataset to 2 components (PC1, PC2), retaining most variance and removing redundancy.
   * Outcome: Enhanced interpretability and algorithm performance.
5. Clustering on Reduced Features:
   * Method: Reapplied K-Means and Hierarchical Clustering using optimized parameters.
   * Outcome: Distinct, well-separated clusters with improved computational efficiency.
6. Evaluation:

Average silhouette score is 0.6617, indicating well-separated clusters.

Graph Explanation:



* X-Axis: Principal Component 1 (PC1).
* Y-Axis: Principal Component 2 (PC2).
* Colour coded labels represent independent clusters

1. Insights of clustering

* Dimensionality reduction using PCA and feature selection has improved clustering results and increased our average silhouette scores.

1. **Anomaly Detection Using Isolation Forest:**

**Isolation Forest code file name:** *MLUL2-Anomaly Detection using ISF.ipynb*

Feature wise analysis:

* + - * Cholesterol (chol):
* Cluster 1: The highest mean cholesterol among anomalies suggests these individuals are likely experiencing anomalies driven by metabolic issues (e.g., hyperlipidemia or cardiovascular risks associated with high cholesterol).
* Cluster 3: The lowest mean cholesterol among anomalies indicates anomalies could stem from other factors, such as elevated blood pressure, rather than cholesterol levels.
  + - * Resting Blood Pressure (trestbps): Cluster 3: High blood pressure is the defining characteristic of anomalies, pointing to potential cardiovascular stress or conditions like hypertension.
      * Heart Rate Achieved (thalch): Cluster 0 & 2: Higher heart rate values indicate that physical exertion or stress may contribute to anomalies, potentially linked to cardiac workload or arrhythmias.
      * ST Depression (oldpeak): Oldpeak is not a critical driver as its values are consistent across clusters hence does not contribute to anomalies

|  |  |
| --- | --- |
| Fig. 1: | Fig. 2 |

***Fig. 1: Feature Variance Across Clusters (Anomalies Only)****: This bar chart highlights the variance of features contributing to anomalies identified by Isolation Forest (ISF), with the highest variance observed in "chol" and "oldpeak."*

***Fig. 2: Anomalies Distribution Across Clusters****: This chart shows the count of normal points and anomalies in each PCA-reduced K-Means cluster, with Cluster 0 containing the highest number of anomalies*.

Cluster-Wise Summary of Anomalies:

* Cluster 0: Moderate cholesterol and resting blood pressure but higher heart rates suggest anomalies due to stress or physical exertion.
* Cluster 1: Main anomaly drivers are high cholesterol levels combined with moderate blood pressure and lower heart rates.
* Cluster 2: Cardiac workload is driven by low cholesterol with elevated heart rates and resting blood pressure
* Cluster 3: Anomalies are due to cardiovascular stress which is indicated by the symptoms such as elevated resting blood pressure with stable heart rates and lower cholesterol.

Interpretation: The anomalies are influenced by distinct feature combinations in each cluster, emphasizing varied underlying causes:

* Cluster 1 anomalies are likely linked to metabolic issues.
* Cluster 3 anomalies suggest hypertensive stress.

1. **Anomaly Detection Using Local Outlier Factor (LOF)**

**LOF Code File Name:** MLUL2 - Clustering & Anomaly Detection using LOF.ipynb

Cluster-Wise Summary of Anomalies:

|  |
| --- |
| ***Fig. 3:*** *The bar chart shows the count of outliers identified in each cluster using the Local Outlier Factor (LOF) method, with Cluster 0 having the highest number of outliers.* |
| ***Fig. 4:*** *The histogram (Fig. 4) displays the distribution of LOF scores across clusters, with a threshold of 1.5 (marked by the red line) separating normal points from outliers.* |

* **Cluster 0**: Inliers are younger with normal metrics and minimal cardiac risks. Outliers are older with slightly lower cholesterol but higher ST depression, suggesting early ischemic risk. **Insight**: Monitor outliers for aging-related ischemia or stress-induced conditions.
* **Cluster 1**: Inliers are older with moderate cholesterol, high BP, and elevated ST depression, indicating hypertension and cardiac risk. Outliers are younger with lower cholesterol and variability, pointing to hidden or early-stage conditions. **Insight**: High-risk group requiring intervention for both inliers and younger individuals.
* **Cluster 2**: Inliers are slightly younger with good heart rates, moderate cholesterol, and low ST depression, suggesting controlled metrics. Outliers are younger with lower BP and cholesterol, possibly indicating early-onset risks. **Insight**: Outliers need attention for subclinical cardiac risks.
* **Cluster 3**: Inliers show moderate BP and cholesterol but high ST depression, indicating ischemic risk. Outliers are older with lower BP and cholesterol but even higher ST depression, suggesting severe ischemia. **Insight**: Outliers represent advanced ischemic or stress-related risks requiring close monitoring.

**Insights from Combined Analysis:**

After implementing anomaly detection on data clusters using both Isolation Forest and Local Outlier factor, we compared the results and insights are given below:

**Isolation Forest vs. LOF:**

* LOF: Detect anomalies by providing extra information by distinguishing inliers and outliers within clusters revealing information on hidden or age-related risk factors.
* Isolation Forest: Detect anomalies by closely matching metabolic issues related to cholesterol and cardiovascular stress related to blood pressure patterns.

**Key Observations:**

* Cholesterol is strong driver of anomalies in both LOF and Isolation Forest hence being important factor in identifying metabolic risk
* Resting BP: Plays a critical role in distinguishing anomalies driven by cardiovascular stress.
* ST Depression (oldpeak): More relevant in LOF clusters (e.g., Cluster 3) for identifying ischemic risks.
* Heart Rate: Irregularities in heart rate could have been due to stress or physical exertion in cluster 3.

**Cluster-Wise Summary achieved from both anomaly detection algorithms:**

* Cluster 1 in Isolation Forest and LOF represents a high-risk group, emphasizing the need for intervention in hypertension and cholesterol management.
* Cluster 3 anomalies identified in both methods show ischemic or stress-related issues, which would require monitoring and stress management programs.
* Younger individuals which were highlighted as outliers in Cluster 0 and Cluster 2 in LOF, highlight the importance of early onset of health risk

**Clinical Implications:**

Based on anomaly detection on our dataset we observed that customized interventions would be needed across all clusters due to different underlying causses such as,

* + Metabolic intervention for high-cholesterol anomalies.
  + Cardiovascular management for elevated BP and ischemic risk.
  + Continuous monitoring of outliers, especially in younger individuals, could help in early detection of hidden risks.

## **Association Rules:**

**Association rules code file name:** *Association\_rules.ipynb*

**Objective**: Through Association rules, we aimed to find correlations and patterns among health metrics and symptoms in individuals with heart disease to enable more reliable diagnosis and risk identification. By analysing combinations of features, we sought to uncover associations linked to the likely existence and severity of heart disease.

**Approach**:

**Data Preprocessing**:

* 1. **Categorical Columns**: One-hot encoded for compatibility with Apriori.
  2. **Numerical Columns**: Missing values and 0’s in ‘*chol’* and ‘*trestbps’* were replaced with NaN for imputation, followed by scaling of all numerical columns using MinMaxScaler.
  3. **Imputation**: KNN imputation was applied to fill missing values, ensuring data integrity.
  4. After-imputation of values, the numerical column scaling was reversed to get the original values, and were binned into appropriate categories (e.g., *age*, *chol*, *thalch*) for suitability with Apriori.
  5. Additionally, the first column of the encoded attributes, which were dropped while one hot encoding, were re-introduced in the dataset based on remaining columns of the same attribute. For example, while one hot encoding sex\_Male was dropped to reduce redundancy. However, it was added back again in this step.

This structured approach enabled the identification of meaningful associations, facilitating insights into the complex nature of heart disease.

**Apriori and association rules implementation**

Dataset filtering by severity: Data was split into subsets based on the num attribute (indicating disease severity), then the models were run and the results were saved as csv files as below:

* Patients with num > 0 (any heart disease).

Results files: *(freq\_items\_num\_1\_4.csv & rules\_num\_1\_4.csv)*

* Patients with num = 1.

Results files: *(freq\_items\_num\_1.csv & rules\_num\_1.csv)*

* Patients with num = 2.

Results files: *(freq\_items\_num\_2.csv & rules\_num\_2.csv)*

* Patients with num = 3.

Results files: *(freq\_items\_num\_3.csv & rules\_num\_3.csv)*

* Patients with num = 4 (severe heart disease).

Results files: *(freq\_items\_num\_4.csv & rules\_num\_4.csv)*

**Findings:**

1. Asymptomatic Chest Pain as a Key Indicator:
   1. cp\_asymptomatic has a strong support value (0.769) for num > 0 in freq\_items\_num\_1\_4, suggesting that individuals with heart disease experience it frequently.
   2. The support for cp\_asymptomatic rises to 0.821 in freq\_items\_num\_4, indicating an even greater occurrence in severe cases.
   3. cp\_asymptomatic is present in several high-confidence rules in rules\_num\_1\_4 and rules\_num\_4, frequently in conjunction with other characteristics like elevated cholesterol and an irregular ECG.
   4. Example Rule (from rules\_num\_4):

Antecedent: {'fbs\_False', 'thalch\_Normal', 'cp\_asymptomatic'}

Consequent: {'sex\_Male'}

Lift: 1.08, Confidence: 1.0

Inference: Particularly in severe cases, asymptomatic chest pain is a major predictor of heart disease, as seen by the high support and confidence levels across datasets.

1. Resting ECG abnormalities:
   1. The support for restecg\_lv hypertrophy in freq\_items\_num\_4 is 0.464, which is significant for extreme instances (num = 4).
   2. LV hypertrophy appears in many rules with high confidence in rules\_num\_4, frequently in conjunction with high cholesterol and angina brought on by activity.
   3. Example Rule (from rules\_num\_4):

Antecedent: {'restecg\_lv hypertrophy', 'cp\_asymptomatic', 'chol\_High'}

Consequent: {'sex\_Male'}

Confidence: 1.0, Lift: 1.08

Inference: LV hypertrophy and ST-T anomalies for advanced coronary artery disease are diagnostically important as shown by their persistent occurrence in severe patients.

1. Exercise-Induced Angina and Oldpeak
   1. In freq\_items\_num\_4, exang\_True (exercise-induced angina) has a support of 0.428, and oldpeak\_high frequently appears in rules with high confidence.
   2. In rules\_num\_1\_4, oldpeak\_high often co-occurs with other risk factors like high cholesterol and abnormal ECG

Inference: Exercise-induced angina and elevated oldpeak values are significant markers of myocardial ischemia and coronary artery disease.

1. Interaction Between Risk Factors
   1. Multiple characteristics, such as elevated cholesterol, irregular ECG, and decreased exercise tolerance, often co-occur with high confidence and lift in rules\_num\_4.
   2. Example Rule (from rules\_num\_4):

Antecedent: {'thalch\_Normal', 'chol\_High', 'cp\_asymptomatic'}

Consequent: {'sex\_Male'}

Confidence: 1.0, Lift: 1.08

Conclusion: This highlights the interplay between these factors in driving severe heart disease.

**Main project findings and managerial implications:**

## Main Project Findings:

1. Clustering and Dimensionality Reduction:

Dimensionality reduction using PCA and feature selection ensured clustering utilizing on non – redundant features. Approach improved clustering results and increased our average silhouette scores.

1. Anomaly Detection Insights:

Isolation Forest and Local Outlier Factor (LOF) revealed distinct patterns in anomalies:

* Cluster 1 anomalies were strongly associated with metabolic issues (elevated cholesterol).
* Cluster 3 anomalies were linked to hypertensive stress (high resting blood pressure) and ischemic issues.
* Younger individuals who are getting detected as outliers in in Clusters 0 and 2 under LOF, have early-onset risk factors, suggesting the importance of proactive health monitoring.

Key risk indicators:

* Elevated cholesterol and resting blood pressure were primary drivers of anomalies.
* Heart rate anomalies were linked to physical exertion and stress.
* In LOF ‘ST depression (oldpeak)’ emerged as a critical factor, specifically for ischemic clusters like Cluster 3.

# **Conclusions and recommendations:**

## Conclusions

1. Clustering by utilising dimensionality reduction (PCA), ensured clusters formed are interpretable and enabled us to uncover precise patterns in heart disease dataset with actionable insights.
2. The combination of anomaly detection and association analysis provided a comprehensive understanding of heart disease risk factors and patterns.
3. Anomalies are driven by distinct feature combinations, emphasizing varied underlying causes (e.g., metabolic issues, cardiovascular stress, ischemia).
4. Association rules revealed critical co-occurrences (e.g., asymptomatic chest pain, high cholesterol, and LV hypertrophy), highlighting their diagnostic significance in severe heart disease.

## Recommendations

1. Implement Personalized Health Interventions:

Focus on addressing specific cluster characteristics:

* + - Cluster 1: Introduce cholesterol-management interventions and promote healthier lifestyles.
    - Cluster 3: Launch programs to target hypertension and ischemia.

1. Promote early screening and preventive care for younger individuals to protect them from emerging risk.
2. Leverage Predictive Analytics:
   * Use findings to develop predictive models for heart disease, integrating key indicators like cholesterol, BP, heart rate, and oldpeak.
   * Deploy machine learning models to classify and predict at-risk individuals based on their feature profiles.
3. Implement Continuous Monitoring Programs:
   * Introduce wearable devices and remote health monitoring to track anomalies in real time, especially for high-risk clusters (e.g., Clusters 1 and 3).
   * Use these programs to detect early signs of disease progression, ensuring timely interventions.
4. Enhance Diagnostic Frameworks:
   * Use association rule insights to refine diagnostic criteria, prioritizing features like asymptomatic chest pain, LV hypertrophy, and oldpeak abnormalities.
   * Train healthcare professionals to recognize key patterns and their interplay in severe heart disease.
5. Invest in Preventative Care:
   * Design community health programs targeting metabolic and cardiovascular risk factors.
   * Educate the public about the importance of managing cholesterol and blood pressure to prevent severe outcomes.