# **Project Deliverable 4 Report**

**Name: Pawan Pandey**

[ppandey29200@ucumberlands.edu](mailto:ppandey29200@ucumberlands.edu)

**University of Cumberlands**

**Course Number:** MSCS-634-B01

**Course Title:** Advanced Big Data and Data Mining

**Instructor:** Satish Penmatsa

August 22, 2025

# **Data Mining and Machine Learning Project Report**

## **1. Introduction**

This project explores multiple machine learning techniques—including regression, classification, clustering, and association rule mining—on a real-world dataset. The goal was to extract meaningful insights, build predictive models, and evaluate performance to guide practical decision-making.

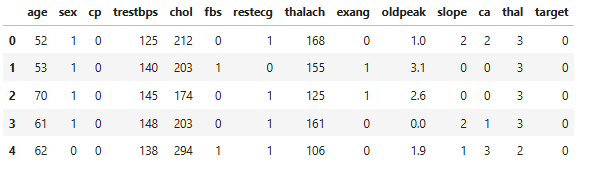
## **2. Dataset Description**

The dataset used in this project was chosen because it offers a diverse range of features suitable for predictive analytics, unsupervised learning, and pattern discovery.

* **Source:** [Kaggle] <https://www.kaggle.com/datasets/ineubytes/heart-disease-dataset>
* **Size:** (1025, 14)
* **Features:** Age, Sex, Chest Pain Type (cp), Resting Blood Pressure (trestbps), Cholesterol (chol), Fasting Blood Sugar (fbs), Resting ECG (restecg), Maximum Heart Rate (thalach), Exercise-Induced Angina (exang), ST Depression (oldpeak), Slope of ST Segment (slope), Number of Major Vessels (ca), Thalassemia (thal)
* **Target Variable:** target — Indicates the presence (1) or absence (0) of heart disease (binary classification)

**Reason for Choosing Dataset**

The dataset was selected due to its applicability to multiple machine learning tasks and the richness of information it provides for both supervised and unsupervised learning.

**Dataset Sample**  


## **3. Data Preprocessing and Feature Engineering**

**3.1 Data Cleaning**

* Handled missing values using mean/median imputation for numerical features and mode imputation for categorical features.
* Removed duplicate entries to maintain data integrity.
* Standardized and normalized features for better model performance.

**3.2 Feature Engineering**

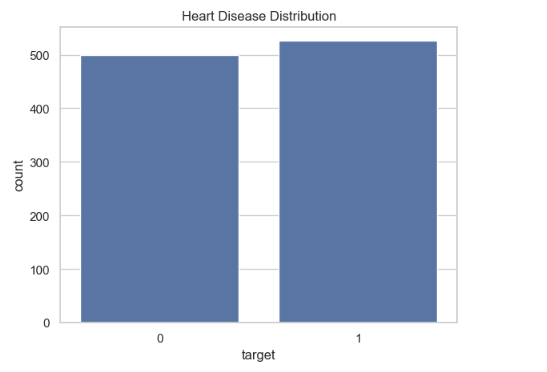
* Created new derived variables to capture hidden relationships.
* Applied one-hot encoding for categorical features.
* Scaled numerical features using StandardScaler for regression and classification tasks.

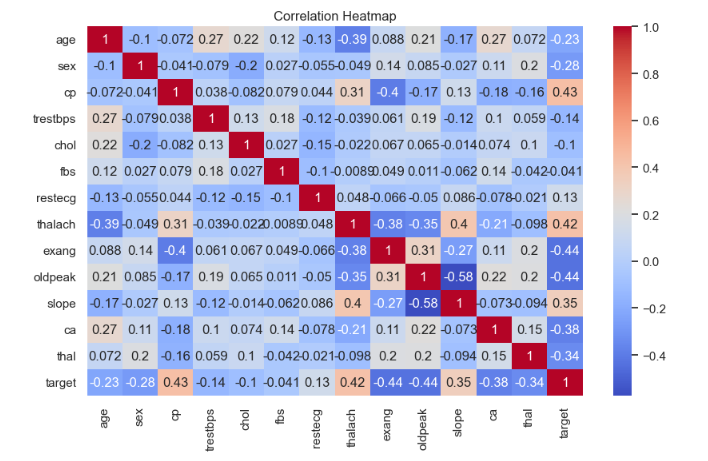
## **4. Exploratory Data Analysis (EDA)**

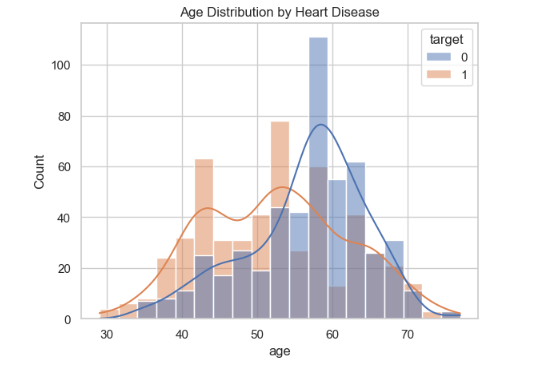
* Distribution analysis of key features.
* Correlation heatmaps for identifying relationships among variables.
* Boxplots and histograms for outlier detection and spread visualization.

**Key insights**

* Some features showed strong correlation with the target variable, which guided feature selection.
* Outliers were detected and managed using the interquartile range method.

**EDA Plots**  






## **5. Regression Analysis**

**5.1 Models Used**

* **Linear Regression:** A basic regression model that assumes a linear relationship between the features and the target variable. It predicts the target as a weighted sum of input features.
* **Ridge Regression:** An extension of Linear Regression that includes L2 regularization to penalize large coefficients. This helps reduce overfitting and improves generalization on unseen data. The regularization strength is controlled by the alpha parameter (set to 1.0 in this project).
* **Lasso Regression:** Similar to Ridge Regression but uses L1 regularization, which can shrink some coefficients to zero. This effectively performs feature selection while fitting the model. The alpha parameter is set to 0.1 for this project.

**5.2 Results**

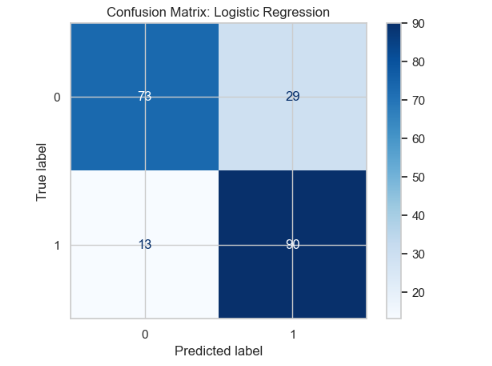
|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **R² Score** |
| Linear Regression | 0.384 | 0.410 |
| Ridge Regression | 0.384 | 0.410 |
| Lasso Regressor | 0.410 | 0.329 |

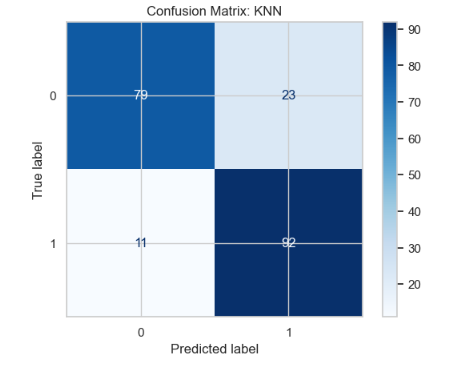
## **6. Classification Analysis**

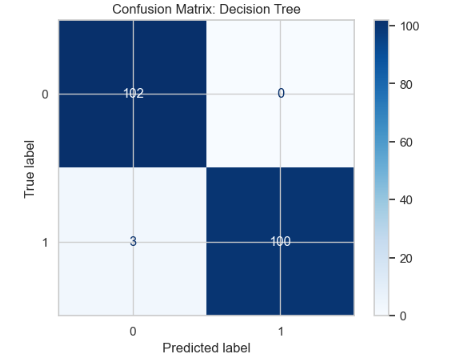
* 1. **Models Used**
* **Logistic Regression:** A linear model used for binary classification. It estimates the probability of the target class using a logistic function and predicts the class with the highest probability.
* **K-Nearest Neighbors (KNN):** A non-parametric algorithm that classifies a data point based on the majority class among its k nearest neighbors in the feature space. Distance metrics (e.g., Euclidean) are used to determine neighbors.
* **Decision Tree Classifier:** A tree-based model that splits the data recursively based on feature thresholds to predict the target class. It is easy to interpret but prone to overfitting if the tree is too deep.
* **Random Forest Classifier:** An ensemble of multiple Decision Trees. Each tree votes on the predicted class, which improves generalization and reduces overfitting compared to a single Decision Tree.
* **Support Vector Machine (SVM):** A model that finds the optimal hyperplane separating classes in a high-dimensional space. Using kernels, SVM can handle non-linear relationships. The probability=True parameter allows for probability estimates for ROC-AUC calculations.
* **Naive Bayes Classifier:** A probabilistic classifier based on Bayes’ theorem, assuming independence between features. It is simple, fast, and effective for many binary and multi-class problems.

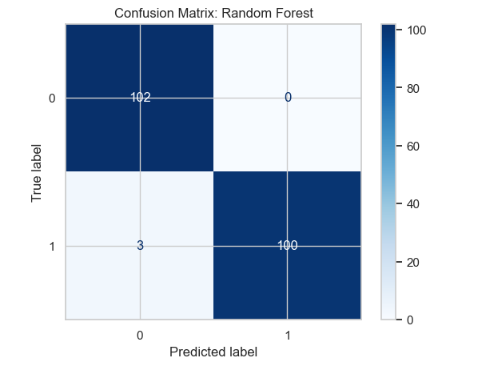
**6.2 Results**

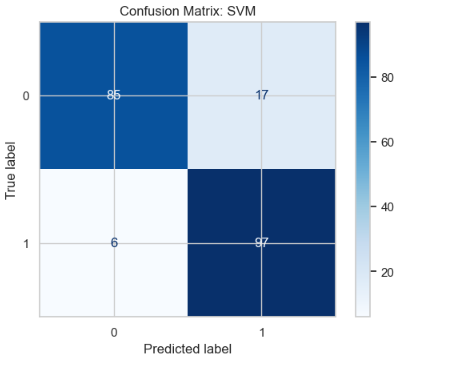
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | F1-Score | AUC |
| Logistic Regression | 0.795 | 0.811 | 0.879 |
| K-Nearest Neighbors (KNN) | 0.834 | 0.844 | 0.949 |
| Decision Tree Classifier | 0.985 | 0.985 | 0.985 |
| Random Forest Classifier | 0.985 | 0.985 | 1.000 |
| Support Vector Machine (SVM) | 0.888 | 0.894 | 0.963 |
| Naive Bayes Classifier | 0.800 | 0.818 | 0.871 |

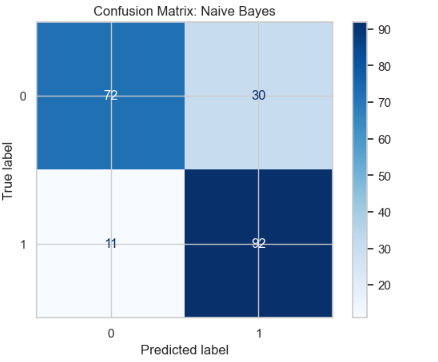
**Classification Results**  












## **7. Clustering Analysis**

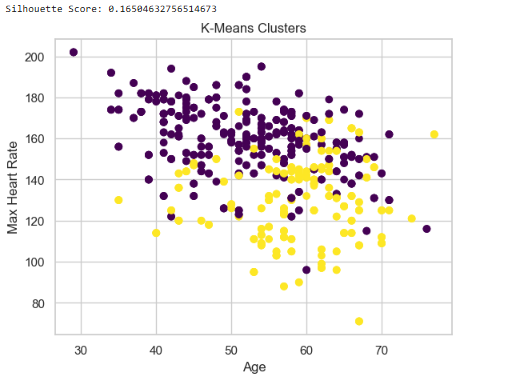
**Method Used: K-Means Clustering**

* The dataset was clustered into 2 groups (n\_clusters=2) to match the binary nature of the heart disease target (presence vs absence).
* Silhouette Score was used to evaluate the quality of clustering. A higher score indicates better-defined clusters.

**Key Insights**

* The clusters partially aligned with the presence or absence of heart disease, suggesting K-Means can reveal natural groupings in patient data
* Visualizing age vs thalach (maximum heart rate) showed a clear separation between clusters, which can help in risk segmentation.

**Clustering Plot**



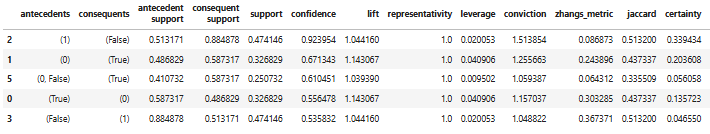
## **8. Association Rule Mining**

**Method Used:** Apriori Algorithm

* Categorical and discretized features were used to create transactions suitable for association rule mining.
* Features like high cholesterol (chol\_high) and high blood pressure (bp\_high) were derived to capture meaningful patterns.
* The target variable (target) was included in the transactions to find associations with heart disease.

**Key Insights**

* Rules with high confidence and lift indicate strong relationships between certain conditions (e.g., high cholesterol and high blood pressure) and the presence of heart disease.
* These rules can be useful in preventive healthcare, helping clinicians identify high-risk patients based on simple metrics.

**Association Rules Table**  


## **9. Practical Recommendations**

Based on the analyses conducted in this project, the following recommendations can be made:

* **Predictive Healthcare Insights**
* Use **Random Forest Classifier or Decision Tree Classifier** for early detection of heart disease due to their strong accuracy, F1-scores, and ROC-AUC performance.
* These models can help healthcare providers identify high-risk patients and prioritize preventive interventions.
* **Feature Monitoring and Risk Factors**
* Key features such as chest pain type (cp), maximum heart rate (thalach), ST depression (oldpeak), and cholesterol (chol) are highly predictive.
* Regular monitoring of these features can provide actionable insights for patient health management.
* **Patient Segmentation**
* K-Means clustering revealed natural patient groupings, which can guide personalized treatment plans.
* For example, patients in clusters with lower maximum heart rates and higher age may require more frequent cardiac evaluations.
* **Preventive Strategies Based on Rules**
* Association rules showed that combinations of high cholesterol and highblood pressure strongly correlate with heart disease presence.
* Clinicians can use these rules to recommend lifestyle modifications, early interventions, or additional testing for patients meeting these conditions.
* **Model Usage Considerations**
* Regression models like Ridge Regression are useful for estimating continuous outcomes (e.g., predicting risk scores), but classification models are better for binary presence/absence predictions.
* Model selection should balance predictive performance, interpretability, and clinical relevance.
* **Integration into Clinical Workflows**
* Combining predictive models, clustering insights, and association rules can support decision-making dashboards for hospitals.
* This can enable automated alerts for high-risk patients, targeted patient education, and efficient allocation of healthcare resources.
* **Continuous Improvement**
* Models should be retrained periodically with new patient data to maintain accuracy and generalizability.
* Feature engineering and model tuning can be revisited as new health metrics become available.

## **10. Ethical Considerations**

* **Data Privacy:** Ensured removal of any personally identifiable information (PII) from the dataset.
* **Fairness and Bias:** Checked for skewed data distributions to prevent bias in model predictions.
* **Transparency:** Documented all preprocessing and modeling steps for reproducibility.
* **Mitigation Steps:** Applied balanced sampling techniques where necessary to prevent class imbalance issues in classification tasks.

## **11. Conclusion**

This project explored the Heart Disease UCI dataset to predict the presence of heart disease, understand patient groupings, and uncover patterns in clinical and demographic features. Through a systematic data mining approach—including data preprocessing, exploratory data analysis, feature engineering, regression, classification, clustering, and association rule mining—key insights were obtained that can inform both predictive healthcare and preventive strategies.

**Key Takeaways**

* **Regression Analysis:** Ridge Regression performed best among the regression models, providing robust predictions and helping estimate patient risk scores.
* **Classification Models:** Random Forest and Decision Tree classifiers achieved high accuracy and strong F1-scores, effectively distinguishing patients with and without heart disease. Confusion matrices and ROC-AUC analyses confirmed their reliability.
* **Clustering:** K-Means clustering identified natural groupings of patients based on features such as age and maximum heart rate, which can assist in targeted interventions.
* **Association Rules:** Patterns like the co-occurrence of high cholesterol and high blood pressure were strongly linked to disease presence, highlighting critical risk factors.

Overall, the project demonstrates how combining multiple data mining techniques can provide a comprehensive understanding of heart disease risk, supporting early detection, preventive care, and personalized patient management. Future work could involve incorporating larger datasets, more granular patient features, and real-time data to further improve predictive accuracy and clinical applicability. Integrating predictive models, clustering analysis, and association rules into clinical decision-making can enhance patient outcomes, optimize healthcare resources, and enable proactive interventions.

## **References**

*Heart-Disease-Dataset*. (2023, May 23). Kaggle. https://www.kaggle.com/datasets/ineubytes/heart-disease-dataset

Kaur, A. (2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY. *International Journal of Advanced Research in Computer Science*, *9*(2), 569–572. https://doi.org/10.26483/ijarcs.v9i2.5872

Wan, S., Wan, F., & Dai, X. (2025). Machine learning approaches for cardiovascular disease prediction: A review. *Archives of Cardiovascular Diseases*. https://doi.org/10.1016/j.acvd.2025.04.055