Heart Disease Dataset Analysis Report

Team Members

• Aisha Samir

• Rola Hany

• Menna Mohsen

• Menna Akram

• Nada Etman

1. Data Preprocessing and Exploration Results

In this phase, the dataset was thoroughly examined to ensure data quality and suitability for modeling. The dataset was checked for missing values and duplicate records. The analysis confirmed that no missing values were present, and all duplicate entries were successfully identified and removed to avoid bias and redundancy in the analysis.

Categorical features were transformed into numerical representations to make them compatible with machine learning algorithms. Binary categorical variables were encoded using 0 and 1 values. Ordinal features, such as **Age Category** and **GenHealth**, were encoded while preserving their natural ordering. Nominal categorical features, including **Race** and **Diabetic**, were converted using one-hot encoding to prevent the introduction of unintended ordinal relationships.

Numerical features such as **BMI**, **Physical Health**, **Mental Health**, and **Sleep Time** were standardized using the **StandardScaler** technique. This step ensured that all numerical variables were on a similar scale, preventing features with larger ranges from dominating the learning process.

Exploratory data analysis was conducted to better understand the dataset. Outliers were visualized using boxplots, allowing for the identification of extreme values. Additionally, the distribution of BMI was analyzed to observe its spread and central tendency. Feature correlation analysis was performed to examine relationships between variables and identify potential multicollinearity.

Finally, the dataset was split into training and testing subsets using an 80%–20% ratio, ensuring sufficient data for model training while retaining a representative portion for evaluation.

Summary

- Missing values: **0**
- Duplicates removed: **Yes**
- Features encoded and scaled
- Outliers and distributions analyzed
- Train-test split: **80% training, 20% testing**

2. Regression Results

A **Linear Regression** model was implemented to predict **Body Mass Index (BMI)** as a continuous target variable. The dataset was divided into 80% training data and 20% testing data to evaluate the model's generalization performance.

The regression model achieved a **Mean Squared Error (MSE)** of **0.893263113785**, an **Root Mean Squared Error (RMSE)** of **0.945125977732**, and an **R^2 score** of **0.1140204559186**.

The RMSE value represents the average magnitude of prediction error in BMI units, indicating that the model's predictions deviate from the actual BMI values by less than one unit on average. The R^2 score suggests that approximately **11.4%** of the variance in BMI can be explained by the selected features. While this indicates limited explanatory power, the model still provides a reasonable baseline for understanding linear relationships within the data and can be improved using more advanced regression techniques.

3. Classification Report – Logistic Regression (Heart Disease)

A **Logistic Regression** model was applied to predict the presence of heart disease (**HeartDisease**). The target variable was converted into binary values, where **No = 0** and **Yes = 1**. Categorical variables were encoded numerically using one-hot encoding, and numerical features were scaled to improve model convergence. Any remaining missing values were handled using mean imputation.

The model was evaluated on the test dataset and achieved an **Accuracy of 0.9138**, indicating that approximately **91%** of predictions were correct.

Evaluation Metrics

- **Precision:** 0.539
  This indicates that among the instances predicted as having heart disease, 53.9% were truly positive.
- **Recall:** 0.100
  This shows that only 10% of actual heart disease cases were correctly identified by the model.
- **False Negatives (FN):** 5033
  These represent patients who actually have heart disease but were incorrectly classified as healthy.

Confusion Matrix

```
[[57889   478]

 [ 5033   559]]
```

A confusion matrix visualization was generated to clearly illustrate the distribution of true positives, true negatives, false positives, and false negatives. This visualization highlights the model's tendency to correctly classify healthy individuals while struggling to identify positive heart disease cases.

4. Precision vs. Recall Trade-Off and False Negatives Analysis

There is a well-known trade-off between **Precision** and **Recall** in classification models. Precision measures how many predicted positive cases are actually positive, while Recall measures how many actual positive cases are successfully identified.

In this healthcare-focused classification task, **False Negatives are particularly critical**. A false negative occurs when a patient with heart disease is predicted as healthy, which may result in delayed diagnosis or lack of necessary medical treatment.

Given the serious consequences of missing heart disease cases, **maximizing Recall is more important than maximizing Precision** in this context. Even if higher Recall leads to more false positives, it ensures that a greater number of patients with heart disease are correctly detected, supporting early intervention and improved healthcare outcomes.

5. Clustering Analysis Summary

In the final stage of the project, **K-Means clustering** was applied to uncover hidden patterns within the heart disease dataset and identify groups of individuals with similar health characteristics. The clustering analysis focused on key lifestyle and health-related features, including **BMI**, **Physical Health**, **Mental Health**, and **Sleep Time**.

Prior to clustering, all selected features were normalized using **StandardScaler** to eliminate scale differences and enhance the performance of the distance-based K-Means algorithm.

The **Elbow Method** was used to determine the optimal number of clusters by examining inertia values across multiple values of K. Based on the elbow point observed in the plot, **K = 3** was selected as the optimal number of clusters. This selection was further supported by the **Silhouette Score**, which indicated a reasonable level of separation and cohesion among clusters.

The final clustering results revealed **three distinct health profiles**, which can be broadly interpreted as **low-risk**, **moderate-risk**, and **higher-risk** groups. These clusters provide meaningful insights into population health patterns and demonstrate the effectiveness of unsupervised learning techniques in healthcare data analysis.

6. Conclusion

This project presented a comprehensive analysis of the Heart Disease dataset using a combination of data preprocessing, exploratory data analysis, supervised learning, and unsupervised learning techniques. The preprocessing stage ensured high data quality through proper encoding of categorical variables, feature scaling, and the removal of duplicate records, providing a solid foundation for reliable modeling and analysis.

Regression analysis using a Linear Regression model demonstrated the ability to predict Body Mass Index (BMI) with a reasonable level of accuracy. Although the R^2 score indicated limited explanatory power, the model served as a useful baseline for understanding linear relationships between health-related features and BMI. Future improvements could include the use of non-linear models or feature selection techniques to enhance predictive performance.

For classification, Logistic Regression was applied to predict the presence of heart disease. While the model achieved high overall accuracy, the evaluation revealed a significant challenge in identifying positive heart disease cases, as reflected by the low recall and high number of false negatives. This highlighted the importance of selecting evaluation metrics carefully in healthcare applications, where maximizing recall is critical to avoid missing patients who require medical attention. Adjusting decision thresholds or applying class balancing techniques could significantly improve the model's clinical usefulness.

Finally, K-Means clustering provided valuable insights into underlying health patterns within the dataset. The identification of three distinct clusters representing different health risk profiles demonstrated the effectiveness of unsupervised learning in population health analysis. These clusters can support targeted interventions and risk stratification strategies in real-world healthcare settings.

Overall, this project illustrates the importance of combining multiple machine learning approaches to gain both predictive power and deeper insight from healthcare data. The results emphasize that model evaluation must align with domain-specific priorities, especially in medical applications where human lives may be affected.