Heart Disease Prediction Using Medical Dataset

Presented by: Vignesh Goswami #2020152,
Anmol Kaw #2021234

1. Overview of the Dataset and Problem Statement

The dataset consists of **13 input features** and **1 target variable** aimed at predicting heart disease presence. The target variable is binary:

- **0**: Healthy (No Heart Disease)
- **1**: Heart Disease

Key Features:

- **Patient Demographics:** Age, Sex
- **Clinical Measurements:** Resting blood pressure (trestbps), Cholesterol (chol)
- **Medical Tests:** Chest Pain Type (cp), Fasting Blood Sugar (fbs), Resting ECG (restecg), Maximum Heart Rate (thalach), Exercise-Induced Angina (exang), ST depression (oldpeak)

Problem Statement: Using this medical dataset, we aim to develop a data processing pipeline and machine learning models to accurately predict the presence of heart disease, enabling early detection and improved patient outcomes.

2. Challenges Faced and Solutions

1. **Missing Values:**
    - **Challenge:** Features like ca and thal contained missing values.
    - **Solution:** Imputed missing numerical values using the median and categorical values using the mode to preserve data consistency.

2. **Categorical Data Encoding:**

- **Challenge:** Features such as sex and cp were categorical and required numerical transformation.
- **Solution:** Binary encoding for sex and one-hot encoding for nominal features like cp.

3. **Imbalanced Dataset:**
    - **Challenge:** Heart disease cases (target = 1) were underrepresented.
    - **Solution:** Applied SMOTE (Synthetic Minority Over-Sampling Technique) to balance the dataset, ensuring equal representation of both classes.

4. **Feature Scaling:**
    - **Challenge:** Features had varying magnitudes, which could affect distance-based models.
    - **Solution:** Used standardization (z-score normalization) to ensure equal contribution of features.

5. **Outlier Detection:**
    - **Challenge:** Extreme values in features like chol and trestbps could bias the model.
    - **Solution:** Removed outliers using the interquartile range (IQR) method.

3. Hypothesis Tests and Conclusions

H1: Relationship Between Age and Heart Disease

- **Null Hypothesis (H0):** Age is independent of heart disease.
- **Alternative Hypothesis (H1):** Age influences heart disease occurrence.
- **Test:** Chi-Square Test of Independence.
- **Conclusion:** The p-value was **< 0.05**, leading us to reject H0. This indicates a significant relationship between age and heart disease.

H2: Gender and Disease Patterns

- **Null Hypothesis (H0):** Disease patterns are identical for men and women.
- **Alternative Hypothesis (H1):** Disease patterns differ by gender.
- **Test:** Independent Samples T-Test.
- **Conclusion:** With a p-value **< 0.05**, H0 was rejected. The results revealed significant differences in symptom patterns between genders.

4. Model Performance: Scaled vs. Unscaled Datasets

Three machine learning models were trained: **Logistic Regression**, **Random Forest**, and **XGBoost**. The models were evaluated on accuracy, precision, recall, and F1-score.

Model	Accuracy (Full Dataset)	Accuracy (Scaled Dataset)	Precision	Recall	F1-Score
Logistic Regression	82%	81%	80%	78%	79%
Random Forest	87%	85%	85%	83%	84%
XGBoost	**89%**	**87%**	**88%**	**87%**	**87%**

Insights:

- **Scaling Impact:** Scaled datasets reduced training time by ~40% but resulted in a slight accuracy drop (~2%).
- **Best Performing Model:** XGBoost achieved the highest accuracy (89%) and best overall performance, followed closely by Random Forest.

5. Conclusion and Future Work

Conclusion:

- XGBoost outperformed all models with an accuracy of **89%**.

- Feature scaling effectively reduced training time but introduced a minor trade-off in accuracy.
- Addressing missing values, outliers, and data imbalance was crucial to improving model robustness and performance.

Future Work:

- Implement advanced feature engineering to uncover hidden patterns and improve predictive accuracy.
- Explore deep learning techniques (e.g., neural networks) to handle more complex patterns.
- Expand the dataset to include diverse demographics and medical conditions for better generalizability.

References

1. **Dataset Source:** UCI Machine Learning Repository: Heart Disease Dataset.
2. **Code Repository:** Link
3. **SMOTE Technique:** Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique.
4. **Chi-Square Test Reference:** Statistics and Probability Tutorials, Khan Academy.
5. **Machine Learning Models:** Pedregosa et al., Scikit-learn: Machine Learning in Python.