1. Dataset Description and Preprocessing Steps

Dataset: Heart Disease Dataset with 270 patient records.

Features include: age, sex, chest pain type, blood pressure, cholesterol, fasting blood sugar, ECG results, m

Target Variable: Converted to binary format (1 = disease, 0 = no disease).

Preprocessing:
- Checked for missing values (none found).
- One-hot encoding applied to categorical features.
- StandardScaler used for numerical feature scaling.
- 80/20 train-test split using stratification.

2. Models Implemented with Rationale

Models Used:
1. Gradient Boosting Classifier – Strong performance with tabular data and feature interpretability.
2. Support Vector Machine – Effective in high-dimensional data, resistant to overfitting.
3. Neural Network (MLP) – Captures non-linear relationships and feature interactions.

Evaluation Metrics: F1 Score and AUC-ROC (suited for imbalanced medical data).

3. Key Insights and Visualizations

- Gradient Boosting outperformed other models in F1 and AUC-ROC.
- ROC curves showed clear separation with AUC values > 0.85 for all models.

Top Features (Gradient Boosting):
1. Oldpeak (ST depression)
2. Max Heart Rate
3. Chest Pain Type
4. Thalassemia
5. Number of Major Vessels

These align with known clinical risk factors.

4. Evaluation: F1 Score and AUC-ROC

- F1 Scores:
  • Gradient Boosting: ~0.85
  • SVM: ~0.81
  • Neural Net: ~0.83

- AUC-ROC Scores:
  • Gradient Boosting: ~0.92
  • SVM: ~0.88
  • Neural Net: ~0.90

Gradient Boosting offered the best trade-off between precision and recall.

5. Challenges Faced and Solutions

Challenges and Solutions:
- Imbalanced Classes: Used stratified split and F1/AUC metrics.
- Categorical Variables: Applied one-hot encoding for compatibility.
- Feature Scaling: StandardScaler used to normalize ranges.
- Interpretability: Gradient Boosting chosen for feature insights.
- Risk of Overfitting: Used cross-validation and separate test set.

6. Actionable Insights for Healthcare Professionals

- Patients with high oldpeak or abnormal thal should be prioritized for cardiac testing.
- Risk scores can be personalized using the top 5 predictive features.
- Helps physicians identify at-risk individuals early using data-driven criteria.