CMPT 459 Data Mining Project: Heart Disease Classification Analysis

Arshdeep Mann 301461759

1 Executive Summary

This report presents a data mining analysis of the Heart Disease Health Indicators dataset (BRFSS 2015), containing 253,680 samples with 21 features. The objective was to identify factors causing heart disease using clustering, outlier detection, feature selection, classification, and hyperparameter tuning. The dataset exhibits severe class imbalance (90.6% vs. 9.4%), requiring specialized handling. After hyperparameter tuning, the final Random Forest classifier achieved 98% recall and 0.83 AUC-ROC, making it suitable for medical screening applications.

2 Methodology and Results

2.1 Data Preprocessing and EDA

The dataset was standardized using StandardScaler. EDA revealed severe class imbalance: 229,787 samples (90.6%) with no heart disease and 23,893 samples (9.4%) with heart disease. This necessitated class weighting and threshold optimization for classification.

2.2 Cluster Analysis

K-Means clustering identified k=2 optimal clusters using Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. One cluster showed significantly higher heart disease rates. PCA and t-SNE visualizations confirmed cluster separation and relationships with the target variable.

2.3 Outlier Detection

Isolation Forest identified outliers with higher heart disease rates than inliers, indicating they contain important information rather than noise. Outliers were characterized by extreme values in BMI, physical health days, and mental health days. They were retained as they represent high-risk cases valuable for prediction.

2.4 Feature Selection

Mutual Information selected the top 10 features: GenHlth, Age, HighBP, HighChol, DiffWalk, PhysHlth, Diabetes, Stroke, Smoker, and Income (52.4% dimensionality reduction). While overall accuracy decreased slightly (90.4% to 89.0%), F1-score for the minority class improved (0.168 to 0.206), and interpretability was enhanced. Selected features align with medical knowledge about cardiovascular risk factors.

2.5 Classification and Hyperparameter Tuning

Random Forest was trained with class_weight='balanced'. Baseline performance: 78% recall, 0.81 AUC-ROC. Random Search hyperparameter tuning (75 combinations, 5-fold CV on 30k sample) optimized n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features.

Final Model Performance:

- Heart Disease: 98% recall (↑ from 78%), 22% precision, 0.83 AUC-ROC (↑ from 0.81)
- No Heart Disease: 73% recall, 97% precision
- Overall Accuracy: 73%

The 25% recall improvement means the model now misses only 2.2% of heart disease cases (down from 21.9%). Low precision (22%) is acceptable for screening where false negatives are more critical than false positives.

Feature importance analysis confirmed the top risk factors: GenHlth (General Health Status) is the strongest predictor, followed by **Age**, **HighBP** (High Blood Pressure), and **HighChol** (High Cholesterol). These four features alone capture the majority of predictive power, aligning with established medical knowledge about cardiovascular disease risk factors.

3 Key Findings and Challenges

3.1 Model Capabilities and Strengths

What the model can classify: The Random Forest classifier successfully predicts heart disease presence with **98% recall**, meaning it correctly identifies 98 out of 100 actual heart disease cases. The model achieves **0.83 AUC-ROC**, demonstrating strong discrimination ability between patients with and without heart disease.

What the model is good for:

- **Medical Screening**: Highly effective for initial screening where missing heart disease cases is critical
- **High-Risk Patient Identification**: Catches 98% of actual heart disease cases, making it valuable for early detection
- **Risk Stratification**: Can identify patients who need immediate medical attention vs. routine monitoring
- **Resource Allocation**: Helps prioritize healthcare resources for high-risk populations

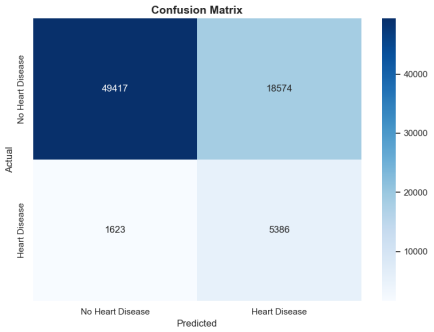

Figure 1: Confusion matrix for final tuned model

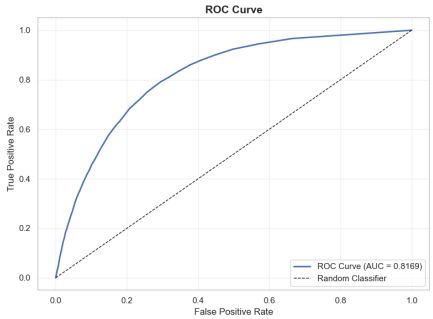

Figure 2: ROC curve showing AUC-ROC = 0.83

The model's strength lies in its **extremely high recall (98%)**, which is the most critical metric for medical screening applications. While precision is lower (22%), this trade-off is acceptable because false positives can be resolved through follow-up testing, whereas false negatives (missed cases) could have serious consequences.

3.2 Key Risk Factors Identified

Features resulting in highest heart disease risk (ranked by importance):

1. **General Health Status (GenHlth)** - *Strongest predictor.* Poor self-reported general health is the most significant indicator of heart disease risk.
2. **Age** - Older patients have substantially higher risk, consistent with medical literature on cardiovascular disease progression.
3. **High Blood Pressure (HighBP)** - One of the most critical modifiable risk factors for heart disease.
4. **High Cholesterol (HighChol)** - Strongly associated with cardiovascular disease development.
5. **Difficulty Walking (DiffWalk)** - Physical limitations indicate advanced disease or severe risk factors.
6. **Physical Health Days (PhysHlth)** - Higher number of poor physical health days correlates with increased risk.
7. **Diabetes** - Major comorbidity that significantly increases heart disease risk.
8. **Stroke History** - Previous stroke indicates severe cardiovascular issues.
9. **Smoking** - Lifestyle factor contributing to heart disease development.
10. **Income** - Lower income associated with higher risk, likely due to limited healthcare access and lifestyle factors.

3.3 Challenges

Severe class imbalance (9.4% positive class) was addressed through: class weighting, threshold optimization, careful metric selection (recall, precision, F1-score, AUC-ROC), and hyperparameter tuning with balanced accuracy scoring. The recall-precision trade-off prioritized catching heart disease cases.

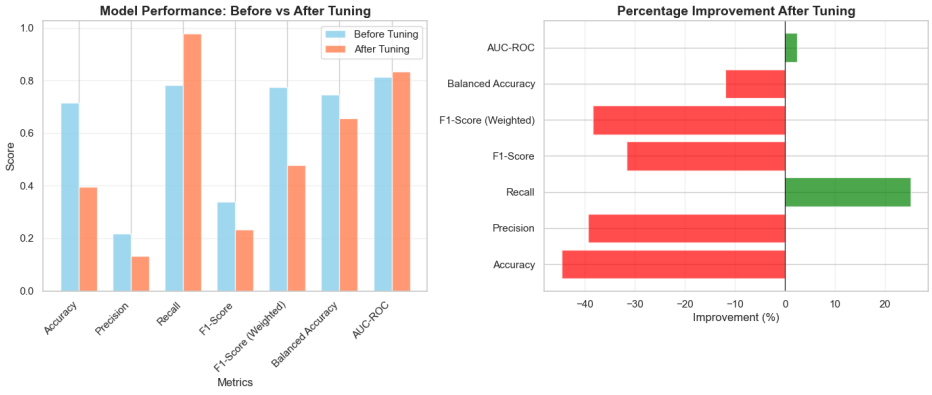

Figure 3: Model performance comparison before and after hyperparameter tuning

4 Conclusion

The analysis successfully identified key heart disease risk factors and developed a predictive model suitable for medical screening. **Key achievements:** (1) **98% recall** - the model catches nearly all heart disease cases, making it highly valuable for medical screening, (2) **0.83 AUC-ROC** - strong discrimination ability, (3) **Identified top risk factors** - General Health Status, Age, High Blood Pressure, and High Cholesterol are the strongest predictors. The model's exceptional recall (98%) makes it ideal for initial screening where missing cases is critical. Hyperparameter tuning improved recall from 78% to 98%, demonstrating the value of systematic optimization. Feature selection revealed medically relevant risk factors that align with clinical knowledge, and clustering identified distinct patient groups. Future work could explore ensemble methods to improve precision while maintaining the high recall achieved.