Q1 Project Report: Heart Disease Predictive Model

Adarsh B, Aryan G
10/21/2024

Table of Contents

1. Project Goal

Project Goal

- Heart Disease Overview: 695,000 people die annually from heart disease in the U.S., which is 1 out of 5 deaths.
- Preventability: 1 in 3 of these deaths could have been prevented with earlier diagnosis.
- Dataset: Combined heart disease dataset from five sources: Hungarian, Cleveland, Switzerland, Long Beach VA, and Statlog (Heart) Data Set.
- Objective: Use the combined dataset to predict if an individual has heart disease based on key medical attributes, contributing to early diagnosis and improved healthcare outcomes.

2. Dataset

Dataset Overview

File: heart.csv

Instances (Rows): 1190

Attributes (Columns): 11 (non-class)

Missing Values:

Few Disguised

Class Attribute: "target"

Class Distribution:

Class 0: No heart disease (47.1%)

Class 1: Heart disease (52.9%)

Dataset Attributes

Age: Age of the patient

Sex: Gender (1 = Male, 0 = Female)

Chest Pain Type: Typical-Angina, Non-Typical-Angina

Resting BP: Resting blood pressure (mmHg)

Cholesterol: Serum cholesterol (mg/dl)

Fasting Blood Sugar: >120 mg/dl (1 = True, 0 = False)

Dataset Attributes Cont.

Resting ECG: Electrocardiographic results (0, 1, 2)

Max Heart Rate: Max heart rate achieved

Exercise Angina: (1 = Yes, 0 = No) (Chest Pain)

Oldpeak: ST depression (Amount HR Decrease After Exercise)

ST Slope: Slope of peak exercise ST segment (Positive, Flat, Negative)

3. Pre-Processing

Pre-Processing

Handling Disguised Entries:

- Replaced cholesterol values of 0 (172 instances) with the median value of 210.
- Replaced resting BP values of 0 (5 instances) with the mean value of 132.

Normalization:

- Min-max normalization for age, BP, cholesterol, max heart rate, oldpeak, ST slope.
- Train-Test Split: 85% training set, 15% testing set, with 1012 training instances and 178 test instances.

4. Selection Algorithms & Model Classifiers

Attribute Selection: Method 1

Method: Correlation analysis

Criteria: Selected attributes with absolute correlation > 0.3

Selected Attributes: ST slope, exercise angina, chest pain type, oldpeak, sex, max heart rate

```
Ranked attributes:
 0.5056  11 ST slope
 0.4815   9 exercise angina
 0.4601   3 chest pain type
 0.3984  10 oldpeak
 0.3113   2 sex
 0.262    1 age
 0.2167   6 fasting blood sugar
 0.1214   4 resting bp s
 0.0731   7 resting ecg
-0.1984   5 cholesterol
-0.4133   8 max heart rate
```

Attribute Selection: Method 2

Method: Cfs subset evaluation

Description: Evaluates attributes based on accuracy and minimizing redundancy through intercorrelation.

Selected Attributes: Age, sex, chest pain type, cholesterol, max heart rate, exercise angina, oldpeak, ST slope

```
Selected attributes: 1,2,3,5,8,9,10,11 : 8
                     age
                     sex
                     chest pain type
                     cholesterol
                     max heart rate
                     exercise angina
                     oldpeak
                     ST slope
```

Attribute Selection: Method 3

Method: Relief Attribute Evaluator

Description: Evaluates attributes by determining how well they separate similar and different class instances. It works well for both discrete and continuous data.

Criteria: Value > 0.04

Selected Attributes: Chest pain type, ST slope, Resting ECG, Fasting blood sugar, Sex, Cholesterol

```
Ranked attributes:
 0.07443    3 chest pain type
 0.06748   11 ST slope
 0.04735    7 resting ecg
 0.04563    6 fasting blood sugar
 0.04395    2 sex
 0.04167    5 cholesterol
 0.0384     1 age
 0.03158    8 max heart rate
 0.02409   10 oldpeak
 0.01692    4 resting bp s
 0.00588    9 exercise angina
```

Attribute Selection: Method 4

Method: Gain Ratio Attribute Evaluator

Description: Measures the predictive power of attributes by using a decision tree and optimizing information gain to handle attributes with many distinct values.

Criteria: Value > .09

Selected Attributes: ST slope, Exercise angina, Chest pain type, Oldpeak, Sex

```
Ranked attributes:
 0.24889  11 ST slope
 0.18536   9 exercise angina
 0.13035   3 chest pain type
 0.10958  10 oldpeak
 0.09122   2 sex
 0.08399   8 max heart rate
 0.05707   5 cholesterol
 0.04702   6 fasting blood sugar
 0.04474   1 age
 0.01967   4 resting bp s
 0.00796   7 resting ecg
```

Attribute Selection: Method 5

Method: Self-Selected Attributes

Description: Manually selected based on underrepresentation in the other 4 and on intuition/familiarity, ensuring important variables are included in the prediction model.

Selected Attributes: Resting ECG, Age, Resting blood pressure, cholesterol, Max heart rate

Classifiers Used

- OneR: We used OneR as a simple baseline to measure the effectiveness of more complex classifiers.

- Random Forest: We chose Random Forest for its versatility and expected it to set a high standard across performance metrics.

- Naive Bayes: We used Naive Bayes to compare the performance of probabilistic and deterministic models.

- J48: We applied J48 to assess how a deterministic decision tree differs from the Random Forest non-deterministic approach.

5. Results

Performance Metrics

- Accuracy: We used accuracy to get an overall view of model performance by pooling all correct predictions.

- TP Rate: We used TP rate to assess how well our models predict both False (No heart disease) and True (Heart disease).

- FP Rate: FP rate helped us measure the percentage of people misdiagnosed as having no heart disease when they actually do.

- AUC: We used AUC as it provides a balance between precision and recall, ideal for our balanced dataset.

- F-Measure: F-measure allowed us to evaluate the balance between precision and recall at the class level for a deeper analysis.

Dataset 1 - Correlation Analysis

Summary Table for Dataset 1:

	Performance Metric					
Classifier Used		Accuracy	TP Rate (False/True)	FP Rate (False/True)	AUC	F-Measure (False/True)
	OneR	.799	.732 .856	.144 .268	.794	.769 .822
	Random Forest	.849	.829 .866	.134 .171	.907	.834 .862
	Naive Bayes	.821	.780 .856	.144 .220	.869	.800 .838
	J48	.832	.817 .845	.155 .183	.870	.817 .845

Dataset 2 - CFS Subset Evaluation

Summary Table for Dataset 2:

	Performance Metric					
Classifier Used		Accuracy	TP Rate (False/True)	FP Rate (False/True	AUC	F-Measure (False/True)
	OneR	.777	.711 .833	.167 .289	.772	.747 .800
	Random Forest	.899	.855 .938	.063 .145	.962	.888 .909
	Naive Bayes	.838	.819 .854	.146 .181	.901	.824 .850
	J48	.872	.843 .896	.104 .157	.882	.859 .882

Dataset 3 - Relief Attribute Evaluator

Summary Table for Dataset 3:

	Performance Metric					
Classifier Used		Accuracy	TP Rate (False/True)	FP Rate (False/True	AUC	F-Measure (False/True)
	OneR	.771	.707 .825	.175 .293	.766	.739 .796
	Random Forest	.950	.927 .969	.031 .073	.970	.944 .954
	Naive Bayes	.855	.841 .866	.134 .159	.914	.841 .866
	J48	.872	.780 .948	.052 .220	.878	.848 .889

Dataset 4 - Gain Ratio Attribute Evaluator

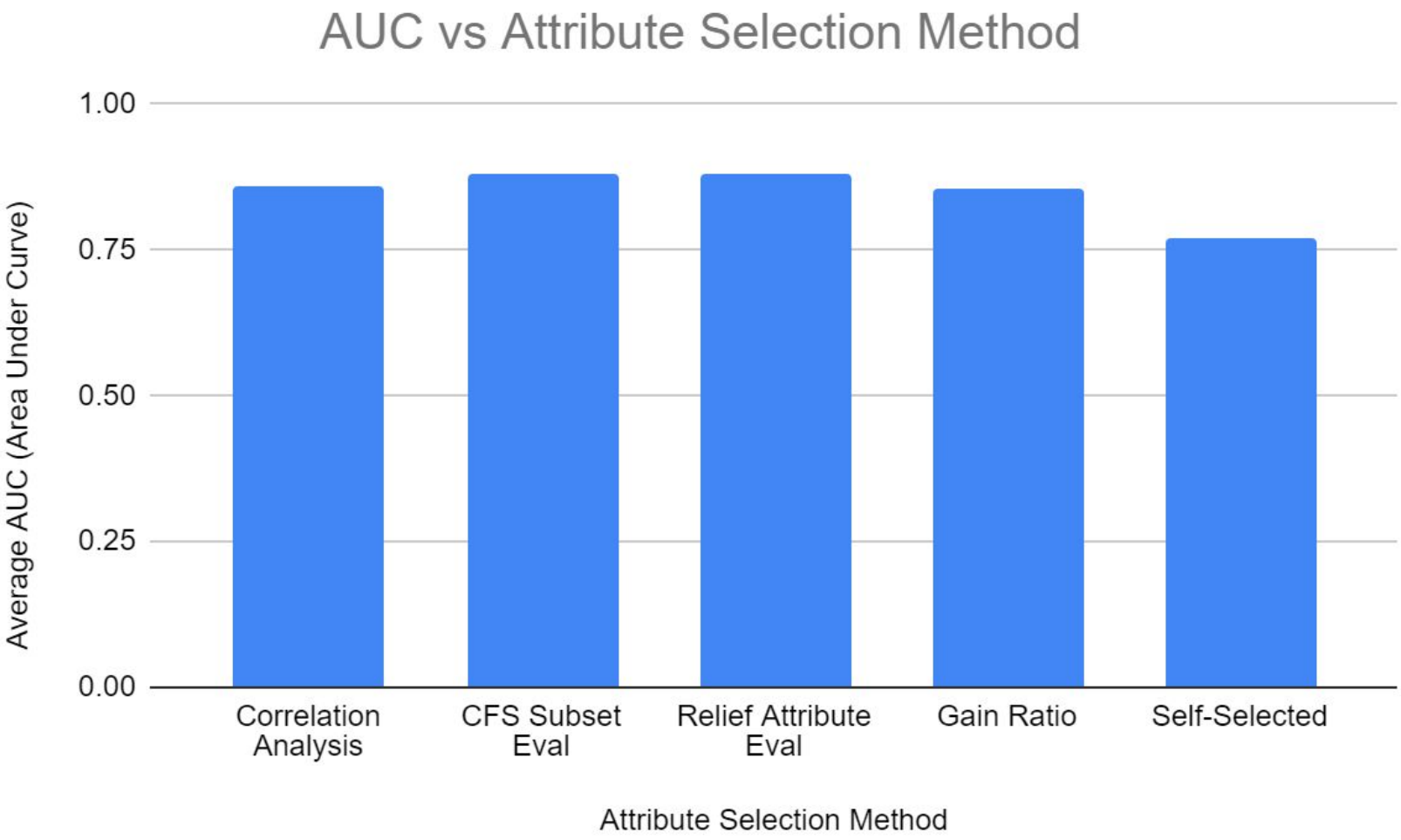
Summary Table for Dataset 4:

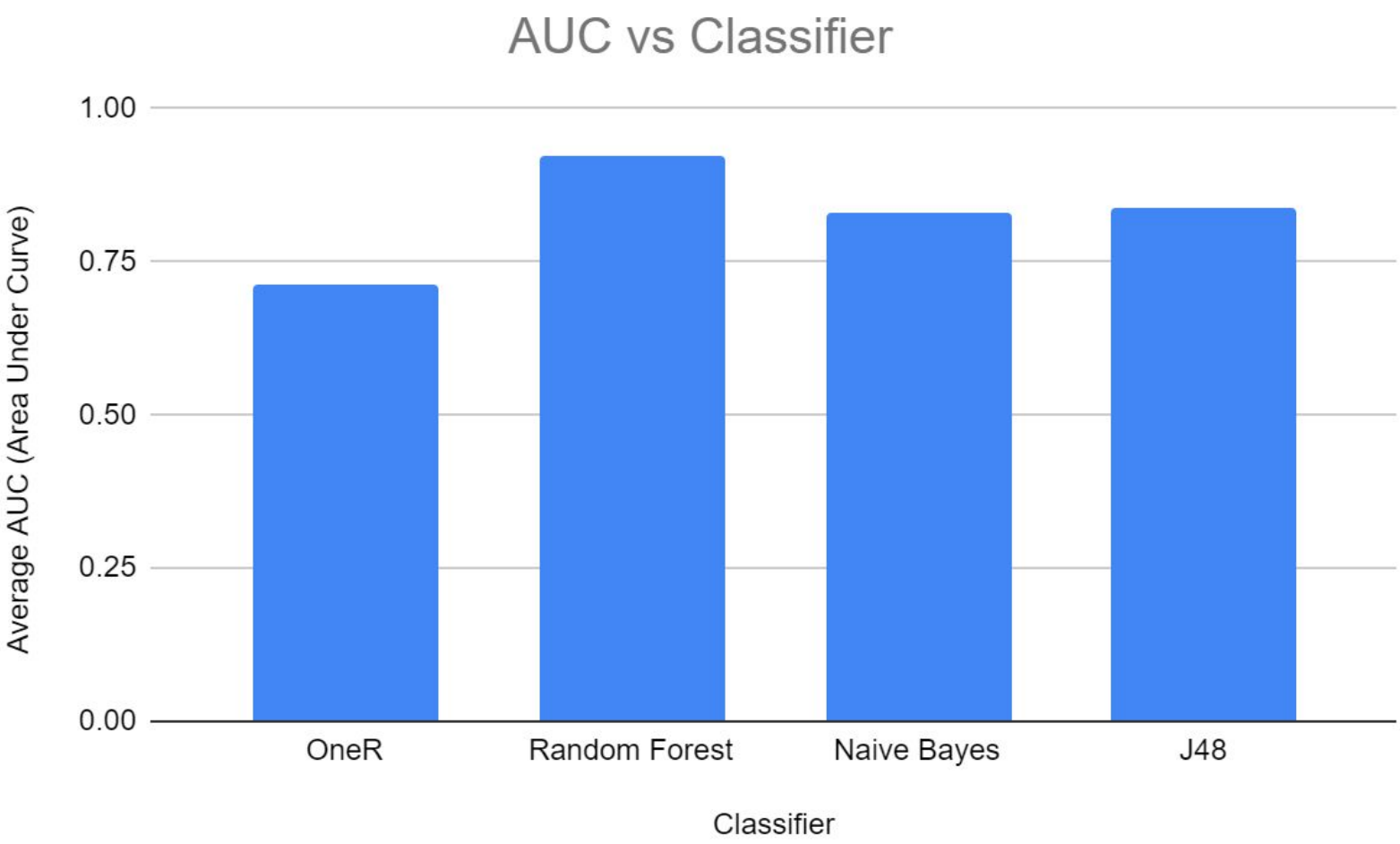
	Performance Metric					
Classifier Used		Accuracy	TP Rate (False/True)	FP Rate (False/True	AUC	F-Measure (False/True)
	OneR	.765	.683 .835	.165 .317	.759	.727 .794
	Random Forest	.866	.878 .856	.144 .122	.930	.857 .874
	Naive Bayes	.754	.634 .856	.144 .366	.866	.703 .790
	J48	.827	.866 .794	.206 .134	.869	.821 .832

Dataset 5 - Self-Selected Attributes

Summary Table for Dataset 5:

	Performance Metric					
Classifier Used		Accuracy	TP Rate (False/True)	FP Rate (False/True)	AUC	F-Measure (False/True)
	OneR	.626	.707 .557	.443 .293	.632	.634 .617
	Random Forest	.849	.817 .876	.124 .183	.903	.832 .863
	Naive Bayes	.698	.683 .711	.289 .317	.760	.675 .719
	J48	.743	.732 .753	.247 .268	.783	.723 .760


AUC vs Attribute Selection Method
1.00
0.75
0.50
0.25
0.00
Average AUC (Area Under Curve)
Correlation Analysis
CFS Subset Eval
Relief Attribute Eval
Gain Ratio
Self-Selected
Attribute Selection Method


AUC vs Classifier
1.00
0.75
0.50
0.25
0.00
Average AUC (Area Under Curve)
OneR
Random Forest
Naive Bayes
J48
Classifier

AUC Per Test

	Attribute Selection Algorithm						
Classifier Used		Correlation Analysis	CFS Subset Eval	Relief Attribute Eval	Gain Ratio	Self-Select	AVG
	OneR	.794	.772	.766	.759	.632	.745
	Random Forest	.907	.962	.970	.930	.903	.934
	Naive Bayes	.869	.901	.914	.866	.760	.862
	J48	.870	.882	.878	.869	.783	.856
	AVG	.860	.879	.882	.856	.770	

6. Analysis

Classifier Performance Overview

- Tested 20 classifiers using 4 models on 5 datasets from different attribute selection methods.
- Random Forest on Dataset 3 (Relief Attribute Eval) achieved the highest AUC of ~97%.
- Random Forest on Dataset 3 also had the highest TP rate for heart disease (96.9%), indicating it accurately identifies heart disease 97% of the time.

Key Comparisons and Findings

- Random Forest and Naive Bayes performed best, with consistent high TP rates, but Random Forest outperformed in challenging datasets.
- In Dataset 5 (self-selected attributes), Random Forest maintained 85% TP rate, while Naive Bayes dropped to 71.1%.
- Naive Bayes struggled with unreliable data, while tree-based models like Random Forest and J48 performed better.

Attribute Selection Insights

- ST Slope was consistently selected by all methods except self-selected, indicating its strong correlation with heart disease.
- Relief Attribute Eval (Method #3) outperformed CFS Subset Eval (Method #2) in both Random Forest and Naive Bayes models.
- Conclusion: Relief Attribute Eval was the best attribute selection method, with Random Forest as the top-performing model.

7. Conclusion

Best Model and Future Improvements

Top Model: Relief Attribute Eval with Random Forest achieved the highest AUC (0.970) and a 96.9% True Positive rate for heart disease.

Consistency: All classifiers, including OneR, had over 75% accuracy, but Random Forest was the most reliable.

Limitations: Using Weka limited us to basic classifiers; for further improvement, we would explore deep neural networks, especially for handling continuous data, as our current models like OneR are not optimized for that.

8. Sources

Sources

Data Source Website:
https://www.kaggle.com/datasets/mexwell/heart-disease-dataset

Files Attached in Google Drive:
arf copies - Pre-processed files used to create models
original csv files - Original data from the source
make_sets.py - Code to split data into 5 datasets with train and test

Sources:
Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

Questions?