Final Project: Comparative Analysis of Machine Learning Models for Heart Disease Prediction Using R

Student Name: Yang Hu

Introduction

- Heart disease remains to be a substantial public health issue.
- Early detection and prevention are critical to reducing its worldwide burden.
- Innovative techniques like PCA and random forest models have improved accuracy in heart disease predictions.
- Predictive algorithms empower healthcare professionals to implement more timely and effective interventions.
- Scientists have made remarkable progress in utilizing data analysis and machine learning.

Problem Statement

Question: How can we accurately predict and prevent heart disease using available health data?

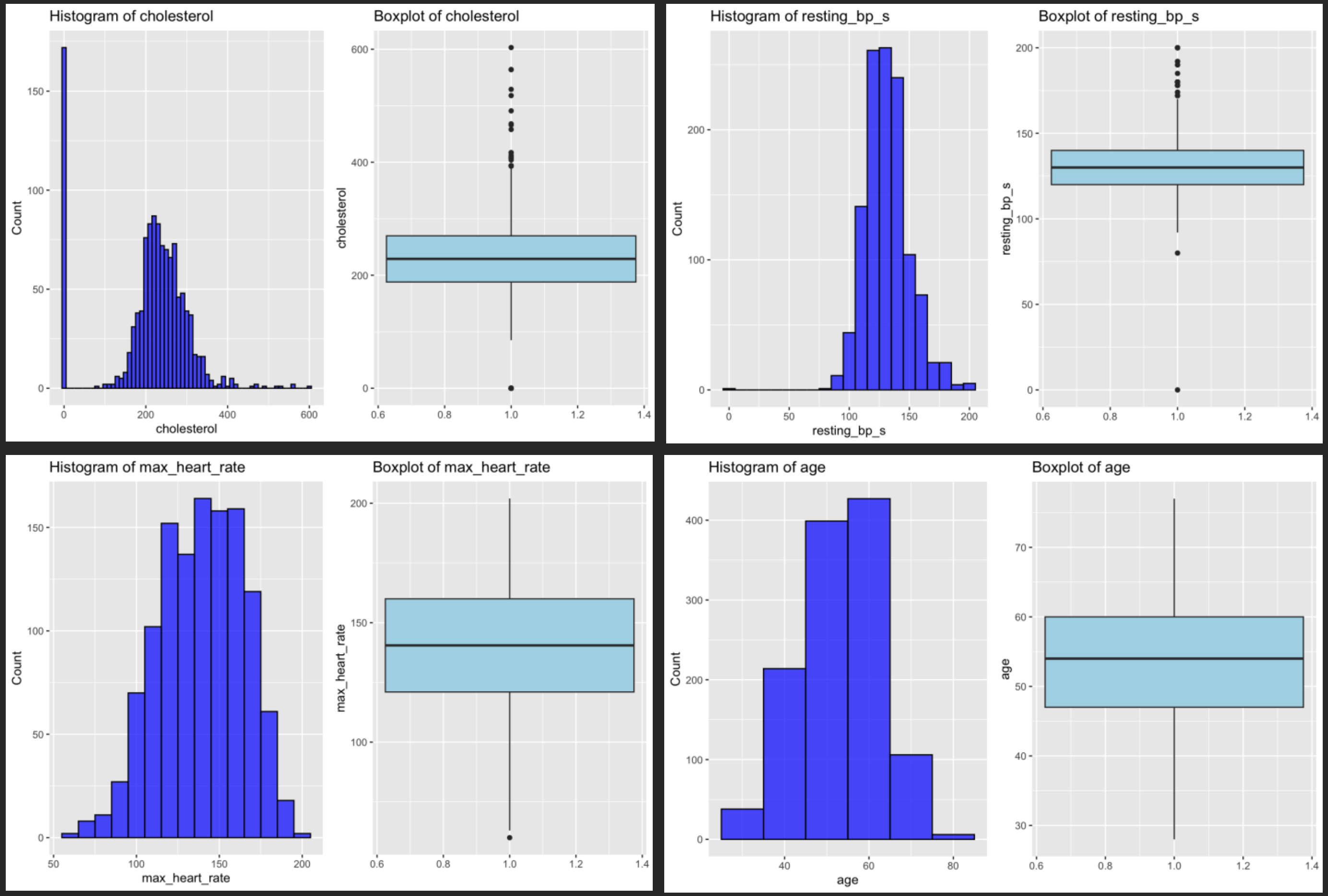
Required Dataset: A comprehensive dataset encompassing various health-related factors

Expected Method: Evaluate various machine learning models to determine which methods accurately predict heart disease risk

Datasets

	age	sex	chest_pain_type	resting_bp_s	cholesterol	fasting_blood_sugar	resting_ecg	max_heart_rate	exercise_angina	oldpeak	ST_slope	target
1	40	1	2	140	289	0	0	172	0	0.0	1	0
2	49	0	3	160	180	0	0	156	0	1.0	2	1
3	37	1	2	130	283	0	1	98	0	0.0	1	0
4	48	0	4	138	214	0	0	108	1	1.5	2	1
5	54	1	3	150	195	0	0	122	0	0.0	1	0
6	39	1	3	120	339	0	0	170	0	0.0	1	0
7	45	0	2	130	237	0	0	170	0	0.0	1	0
8	54	1	2	110	208	0	0	142	0	0.0	1	0
9	37	1	4	140	207	0	0	130	1	1.5	2	1
10	48	0	2	120	284	0	0	120	0	0.0	1	0
11	37	0	3	130	211	0	0	142	0	0.0	1	0
12	58	1	2	136	164	0	1	99	1	2.0	2	1
13	39	1	2	120	204	0	0	145	0	0.0	1	0
14	49	1	4	140	234	0	0	140	1	1.0	2	1
15	42	0	3	115	211	0	1	137	0	0.0	1	0
16	54	0	2	120	273	0	0	150	0	1.5	2	0
17	38	1	4	110	196	0	0	166	0	0.0	2	1
18	43	0	2	120	201	0	0	165	0	0.0	1	0
19	60	1	4	100	248	0	0	125	0	1.0	2	1
20	36	1	2	120	267	0	0	160	0	3.0	2	1
21	43	0	1	100	223	0	0	142	0	0.0	1	0
22	44	1	2	120	184	0	0	142	0	1.0	2	0
23	49	0	2	124	201	0	0	164	0	0.0	1	0
24	44	1	2	150	288	0	0	150	1	3.0	2	1
25	40	1	3	130	215	0	0	138	0	0.0	1	0
26	36	1	3	130	209	0	0	178	0	0.0	1	0
27	53	1	4	124	260	0	1	112	1	3.0	2	0
28	52	1	2	120	284	0	0	118	0	0.0	1	0
29	53	0	2	113	468	0	0	127	0	0.0	1	0

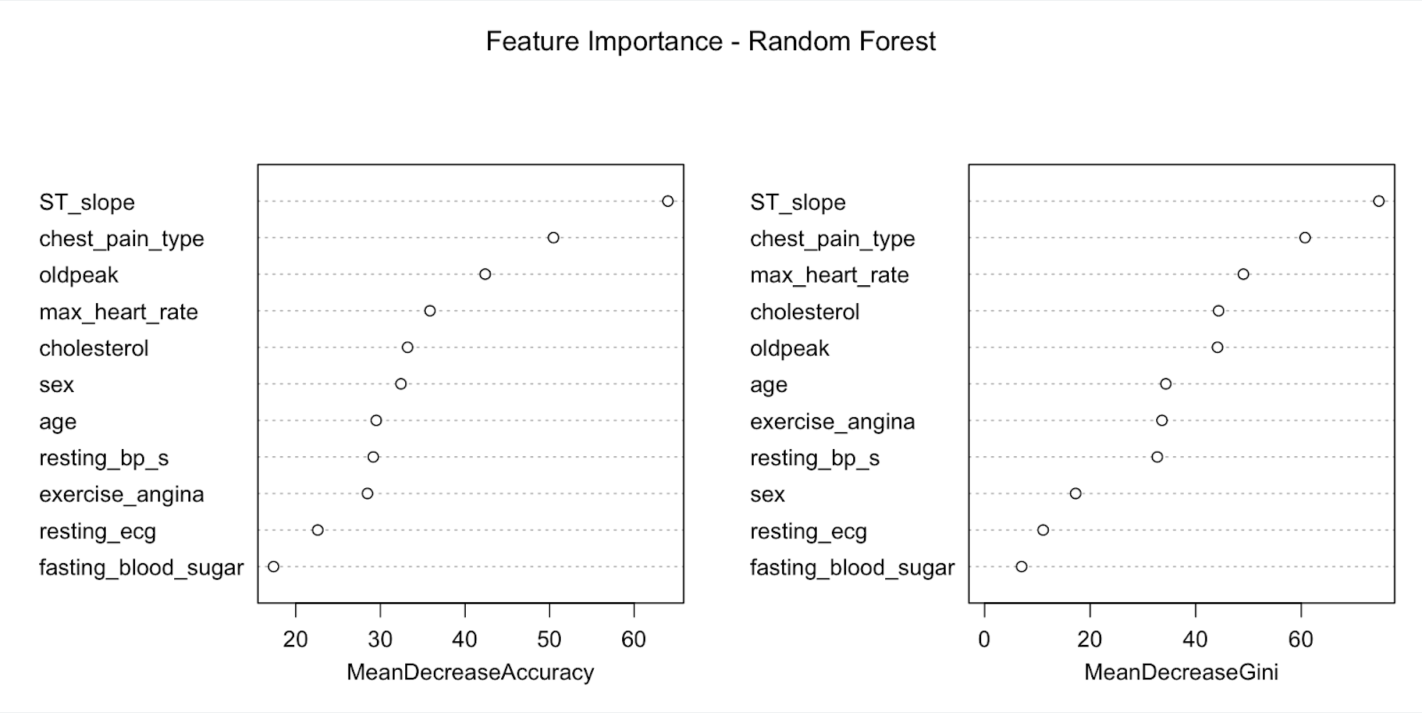
- Heart Disease Dataset (Siddhartha, 2024)
- This dataset *merges data from five different sources, including Cleveland, Hungarian, Switzerland, Long Beach VA, Statlog (Heart) Data Set, into a single, comprehensive dataset with 1,190 instances and 11 features. It's well-suited for general heart disease prediction and multivariate analysis.*



Analysis

- Age Distribution: Most individuals are aged 40-65, peaking in the late 50s to early 60s, with a slight skew towards younger ages. Outliers are present, but the majority are middle-aged.
- Resting Blood Pressure: The distribution centers around 120-140 mm Hg, with a median slightly above 130 mm Hg. Most individuals have normal to slightly elevated blood pressure.
- Cholesterol: The distribution is skewed, with most values around 200 mg/dL and a median of 230 mg/dL, suggesting elevated cholesterol levels. The presence of high outliers and potential data entry issues (spike at 0 mg/dL) is noted.
- Maximum Heart Rate: The distribution peaks around 150 bpm, with a median slightly above this value.

Feature Importance



- Key Predictors: ST Slope, chest pain type, and maximum heart rate are the most critical features in predicting heart disease, significantly influencing model accuracy.
- Secondary Factors: Cholesterol, age, and resting blood pressure also contribute but are less impactful compared to the top predictors.
- Minor Impact: Fasting blood sugar and resting ECG have minimal influence, indicating they play a secondary role in this heart disease prediction model.

Model Comparison Standards

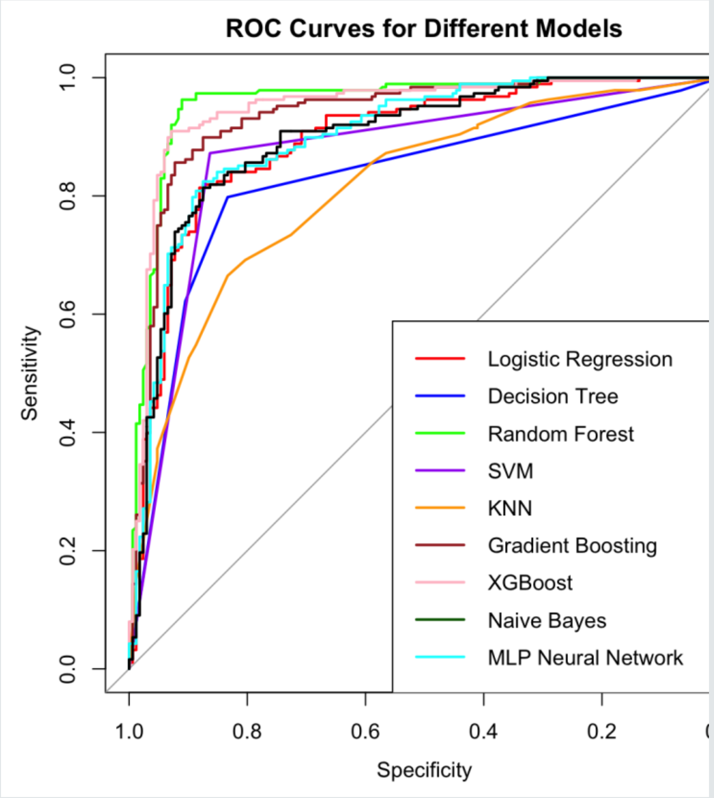
AUC (AREA UNDER THE CURVE): MEASURES OVERALL MODEL PERFORMANCE BY EVALUATING HOW WELL THE MODEL DISTINGUISHES BETWEEN CLASSES. A HIGHER AUC INDICATES BETTER PREDICTIVE ACCURACY.

F1-SCORE: BALANCES PRECISION AND RECALL, PROVIDING A SINGLE METRIC THAT REFLECTS BOTH FALSE POSITIVES AND FALSE NEGATIVES. USEFUL FOR IMBALANCED DATASETS WHERE ACCURACY ALONE IS INSUFFICIENT.

FALSE NEGATIVE RATE (FNR): CRUCIAL FOR HEART DISEASE PREDICTION, AS FALSE NEGATIVES CAN LEAD TO MISSED DIAGNOSES. LOWER FNR IS PRIORITIZED TO ENSURE AT-RISK INDIVIDUALS ARE CORRECTLY IDENTIFIED FOR PREVENTIVE MEASURES.

Model Comparison

- Random Forest: Best model for heart disease prediction, with the highest AUC and F1-score, low false negatives, and strong feature importance insights. Ideal for complex, non-linear data due to its ensemble approach.
- XGBoost: Highly effective with the lowest false negative rate, excellent for large datasets with complex patterns. Requires careful tuning to prevent overfitting and maximize accuracy in medical diagnostics.
- Other Models: Gradient Boosting, SVM, Naive Bayes, and Logistic Regression showed moderate performance, each with specific strengths and limitations. KNN and Decision Tree underperformed, with KNN having potential for improvement through optimization.



Conclusion

- **Top Models**: Random Forest and XGBoost outperformed other models, with the highest AUC and F1 scores, making them ideal for predicting heart disease by effectively capturing complex patterns.
- **Baseline and Other Models**: Logistic Regression and Naive Bayes provided reasonable baselines, while SVM and Gradient Boosting also performed well but slightly less effectively. Decision Tree and KNN had lower performance but could improve with hyperparameter tuning.
- **Feature Importance**: Random Forest's feature importance analysis highlighted key variables driving predictions, offering valuable insights for further exploration and optimization.

Reference

Lin, Z., Chen, S., & Chen, J. (2023). Exploring heart disease prediction through machine learning techniques. *EITCE '23: Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*, 964–969. https://doi.org/10.1145/3650400.3650563

Siddhartha, M. (2024, April 8). *Heart disease dataset*. Kaggle. https://www.kaggle.com/datasets/mexwell/heart-disease-dataset/

Wang, Y. (2023). Research on Predictive Algorithms for cardiovascular Disease. *ISAIMS '23: Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, 304–314. https://doi.org/10.1145/3644116.3644169

Wang, Z., & Tan, X. (2022). Application of double sensitive cost random forest in heart disease detection. *ISAIMS '22: Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*, 559–562. https://doi.org/10.1145/3570773.3570867