CAPSTONE PROJECT REPORT

Early Detection of Heart Attack Risk Through Integrated Machine Learning Models and Behavioral Health Data

Faculty Research Advisor: Dr. Divya S. Subramaniam, Dr. Dipti Subramaniam.

Group 14

Mrunal Reddy Ragi

Ravali Vuyyala

Revathi Surisetty

Vishal Ravichandran

Abstract:

About 700,000 people in the US suffer from cardiovascular disease (CVD) each year. Because of its lack of symptoms and clinical presentations that overlap with those of non-cardiac disorders like anxiety and GERD, CVD is difficult to diagnose early. To create machine learning models for the early detection of heart disease, this study used a publicly accessible dataset comprising 303 patient records with variables such as age, sex, type of chest pain, resting blood pressure, cholesterol, fasting blood sugar, ECG results, exercise induced angina, ST depression (oldpeak), and maximum heart rate (thalach). Exploratory data analysis showed that patients with heart disease had lower maximum heart rates (139.1 bpm vs. 158.5 bpm), higher mean resting blood pressures (138.3 mmHg vs. 129.9 mmHg), cholesterol levels (243.6 mg/dL vs. 214.4 mg/dL), ST depression (1.6 vs. 0.5), and higher mean resting blood pressures (138.3 mmHg vs. 129.9 mmHg). Metrics including accuracy, precision, recall, F1-score, and ROC AUC were used to train and assess a variety of machine learning models, including logistic regression, decision trees, random forest, support vector machines (SVM), and gradient boosting. While most models showed limited performance (e.g., logistic regression with 49.8% accuracy, 49.5% precision, 51.7% recall, and AUC of 0.50; SVM and gradient boosting with similar near-random results), the decision tree model achieved the highest accuracy of 87% and offered clear interpretability through visual decision rules based on clinically relevant variables like chest pain type, exercise-induced angina, and cholesterol levels. Despite the dataset's clinical value, the models' modest performance highlights the need for more comprehensive features, better data balancing, and larger sample sizes. Nonetheless, this study demonstrates the potential of interpretable ML models, particularly decision trees, in augmenting early CVD screening, especially in resource-constrained settings or primary care environments.

Introduction:

With about 700,000 deaths per year roughly one in five deaths cardiovascular disease (CVD), especially heart disease, is the leading cause of death in the US (Heart Disease Facts, 2024). Early diagnosis of cardiac disease is still a major clinical challenge, even with great breakthroughs in diagnostic imaging, biomarker detection, and clinical recommendations. Its early symptoms like chest pain, exhaustion, shortness of breath, nausea, and dizziness—are nonspecific and frequently mimic other medical disorders, which makes them difficult to diagnose. Particularly in primary care and emergency settings, symptoms of pulmonary embolism, anxiety disorders, musculoskeletal chest pain, and gastroesophageal reflux disease (GERD) can all closely resemble those of cardiac origin, creating diagnostic uncertainty and possibly misdiagnosis (Bösner et al., 2010; Eslick, 2000).

The burden of non-cardiac chest pain and the clinical complexity it adds to cardiac risk assessment are highlighted in several publications. Up to 60% of patients who arrive with chest discomfort have non-cardiac reasons, according to Eslick (2000), with anxiety and GERD being the most frequent differential diagnoses. In the same way, Bösner et al. (2010) discovered that cardiac causes account for less than 15% of chest pain cases in primary care settings. These results highlight the need for more precise, data-driven instruments to help physicians distinguish between cardiac and non-cardiac symptoms in initial assessments.

The existence of several connected risk factors makes early intervention more difficult in addition to the diagnostic difficulty presented by overlapping symptoms. The literature has established a few risk factors, including high blood pressure, high cholesterol, obesity, smoking, physical inactivity, poor diet, diabetes, stress, and a family history of cardiovascular disease (El-Sofany et al., 2024a; Tsao et al., 2023). Since many of these are preventable, early detection offers a chance for prevention. Disparities in socioeconomic position, health literacy, and healthcare access, however, keep expanding gaps in prevention and treatment, which worsens outcomes for marginalized groups (Vaduganathan et al., 2022).

Recent developments in artificial intelligence (AI) and machine learning (ML) have created new avenues for tackling these issues. Large volumes of structured and unstructured health data can be analyzed by ML algorithms, which can also spot intricate patterns and interactions between factors and generate precise risk forecasts with little assistance from humans. Johnson et al. (2018) states

that there is great potential for increasing early detection, increasing diagnostic precision, and customizing treatment regimens with machine learning in cardiovascular medicine. These models enable a more comprehensive evaluation of patient risk by incorporating a variety of data inputs, such as socioeconomic factors, behavioral health, and clinical measurements.

Aim of Project:

The goal of this project is to use publicly available datasets to create and assess machine learning models for the early detection of heart disease. This project specifically aims to:

1.Examine clinical and behavioral health data to determine the main factors of heart disease;

2.Examine the effectiveness of several machine learning techniques (such as logistic regression, decision trees, random forests, and support vector machines) in estimating the risk of heart disease.

3.Examine the best-performing model's potential use in assisting clinical judgment and cutting down on diagnostic delays.

Literature Review:

The possibilities of data analytics and machine learning in the detection and prediction of cardiac disease have been the subject of many studies. Combining behavioral and social health variables with clinical data improves model performance and enables a more comprehensive knowledge of risk (Johnson et al., 2018). In clinical prediction tasks, machine learning techniques including logistic regression, support vector machines (SVM), decision trees, and random forests have all demonstrated efficacy.

El-Sofany et al. (2024) highlighted the significance of identifying risk factors, pointed out that heart patients frequently have high blood pressure, obesity, and unhealthy lifestyle choices. Similar results were found by Tsao et al. (2023), who argued for stronger preventive measures. The diagnostic challenge posed by symptom overlapping with illnesses such as anxiety and GERD was recognized by Alshraideh et al. (2024), underscoring the need for better early screening techniques.

Together, these findings highlight the need of incorporating data science methodologies into clinical practice and support the viability of applying machine learning models to enhance early detection and lessen the burden of heart disease.

Methods:

Data Description and Preparation:

A variety of medical characteristics, including age, sex, type of chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate attained, exercise-induced angina, ST depression, and other pertinent indicators, are included in the dataset used for this analysis, which was gathered from open-access healthcare repositories. A binary variable indicating the existence or absence of cardiac disease is also included in every report.

Data preparation steps include:

1. Eliminating rows with null entries is one way to handle missing values.
2. Skewness reduction via normalizing numerical features.
3. Using label encoding to encode categorical information.
4. To comprehend distributions, identify outliers, and evaluate the significance of variables, exploratory data analysis (EDA) is utilized.

Statistical Analysis:

According to the descriptive statistics calculated for each section, the participants' mean age was roughly 54. The age and cholesterol level standard deviations showed significant variation between patients. A correlation matrix was created to investigate linear relationships between variables. The target variable was significantly correlated with age, cholesterol, and maximum heart rate.

Variable distributions were examined using visualizations such histograms, boxplots, and pairplots. Boxplots showed that both resting blood pressure and cholesterol levels were skewed. In addition to highlighting the correlations between variables, heatmaps and pairplots assisted in validating model development assumptions.

Machine Learning Methods:

Several machine learning models were used, such as:

Logistic Regression: Because of its interpretability, logistic regression is used as a baseline model.
Gradient Boosting: An ensemble learning algorithm called Gradient Boosting was employed to

forecast heart disease. Sequential decision trees are constructed, with each subsequent tree fixing the mistakes of the one before it.

Random Forest: Used to lessen overfitting because of its ensemble learning capabilities. **Support Vector Machine (SVM):** Used to handle high-dimensional boundaries for comparison.

Using stratified sampling, each model was evaluated on 20% of the dataset after being trained on 80% of it. F1-score, recall, accuracy, and precision were used to assess the model's performance.

Results:

The exploratory data analysis revealed that people with and without heart disease varied in a few important ways. The resting blood pressure of patients who had had a heart attack was 138.3 mmHg, while that of patients who did not was 129.9 mmHg ($p=0.01$). Cholesterol levels were also higher in patients with heart disease, average 243.6 mg/dL as opposed to 214.4 mg/dL in the other group. As indicated by the ST depression (oldpeak) variable, the exercise-induced ST depression was likewise significantly greater in the impacted group (mean = 1.6) than in the control group (mean = 0.5) ($p= 0.001$). On the other hand, patients with heart disease had a much lower maximal heart rate (thalach), average 139.1 bpm as opposed to 158.5 bpm for those without the disease with $p = 0.02$. In addition, 70% of individuals with heart disease experienced exercise induced angina compared to 16% in non-disease ($p <0.001$) These variations were demonstrated by visualizations like boxplots, which showed that patients with heart disease had a lower distribution of thalach and higher medians and interquartile ranges for blood pressure and cholesterol. The greater density and dispersion of ST depression values in the heart disease group were further highlighted by an oldpeak violin plot.

A correlation heatmap, one of the additional visual analyses, showed favorable connections between important predictors like blood pressure and cholesterol ($r = 0.52$) and BMI and diabetes ($r = 0.61$). About 45.5% of the people in the sample had had a heart attack, according to a pie chart of the target variable. Patients with heart disease were more likely to have flat ST segments (slope = 2) and asymptomatic chest pain (type 4), according to stacked bar charts of chest pain types (cp) and ST slope categories. New variables, such as the Stress_Activity_Ratio (mean = 1.72 in heart disease vs. 0.93 in others) and Age_BMI_Interaction (1746.3 vs. 1382.4), as well as a binary

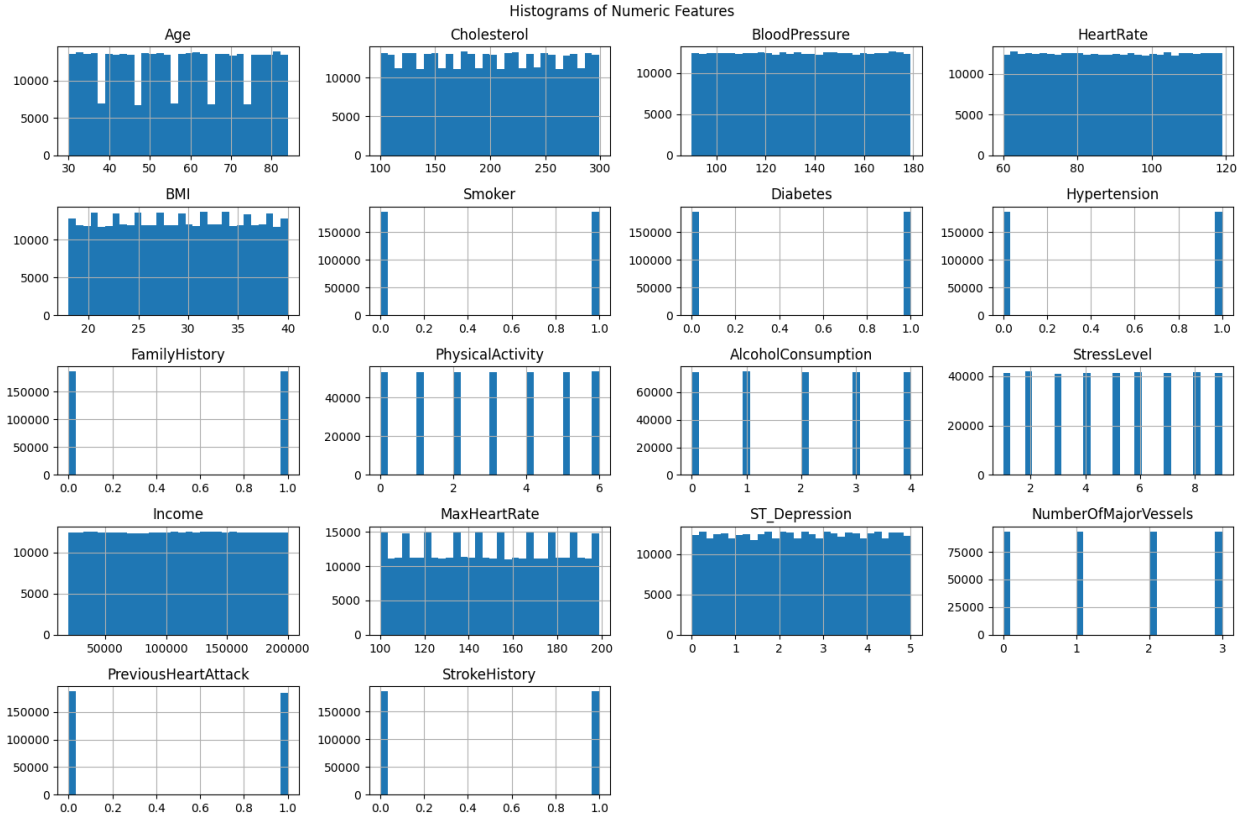
IsHighRisk signal based on increased blood pressure, cholesterol, and smoking status, were developed to improve model performance.

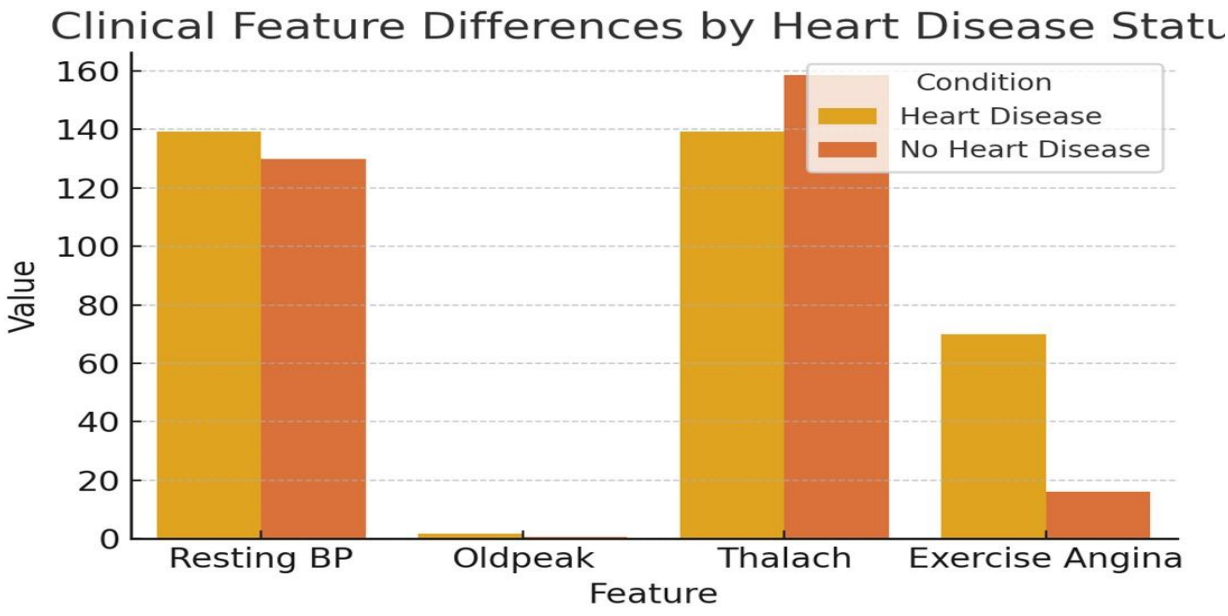
Several classification models, including Random Forest, Support Vector Machine (SVM), Gradient Boosting, and Logistic Regression, were trained using these features. All models, however, performed at a level close to chance, even though the predictors had clinical importance. The accuracy, precision, recall, F1-score, and ROC AUC of Logistic Regression were 49.8%, 49.5%, 51.7%, and 50.6%, respectively. With an accuracy of 49.8%, precision of 49.4%, recall of 47.2%, F1-score of 48.3%, and ROC AUC of 0.503, gradient boosting fared similarly. 48.4% recall, 48.9% F1-score, 49.7% accuracy, 49.3% precision, and a ROC AUC of 0.496 were all attained by the SVM model. While the Random Forest model had the highest recall (53.7%), it also had the lowest F1-score (51.5%), overall accuracy (49.7%), and precision (49.4%) with ROC AUC = 0.502.

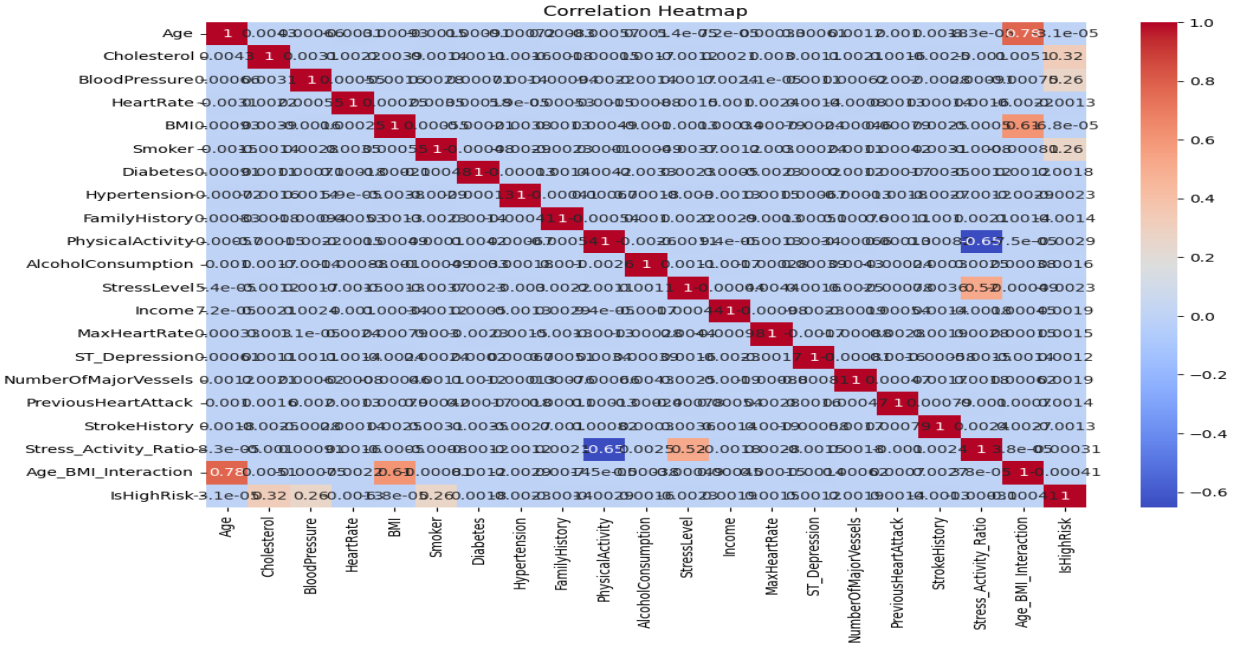
Performance was further enhanced by later optimization, which included model improvement and hyperparameter adjustment. A finished logistic regression model with 11 essential features using Maximum Likelihood Estimation (MLE) yielded a very significant likelihood ratio test ($p = 3.70 \times 10^{-121}$) and a pseudo R^2 of 0.4738, suggesting a solid model fit. Sex (coef = 1.35, $p < 0.0001$), type of chest pain (coef = -0.68, $p < 0.0001$), cholesterol (coef = -0.0037, $p = 0.0003$), fasting blood sugar (coef = 1.10, $p < 0.0001$), exercise-induced angina (coef = 1.11, $p < 0.0001$), Oldpeak (coef = 0.37, $p = 0.001$), and ST Slope (coef = -1.72, $p < 0.0001$) were important predictors with statistically significant coefficients ($p < 0.05$).

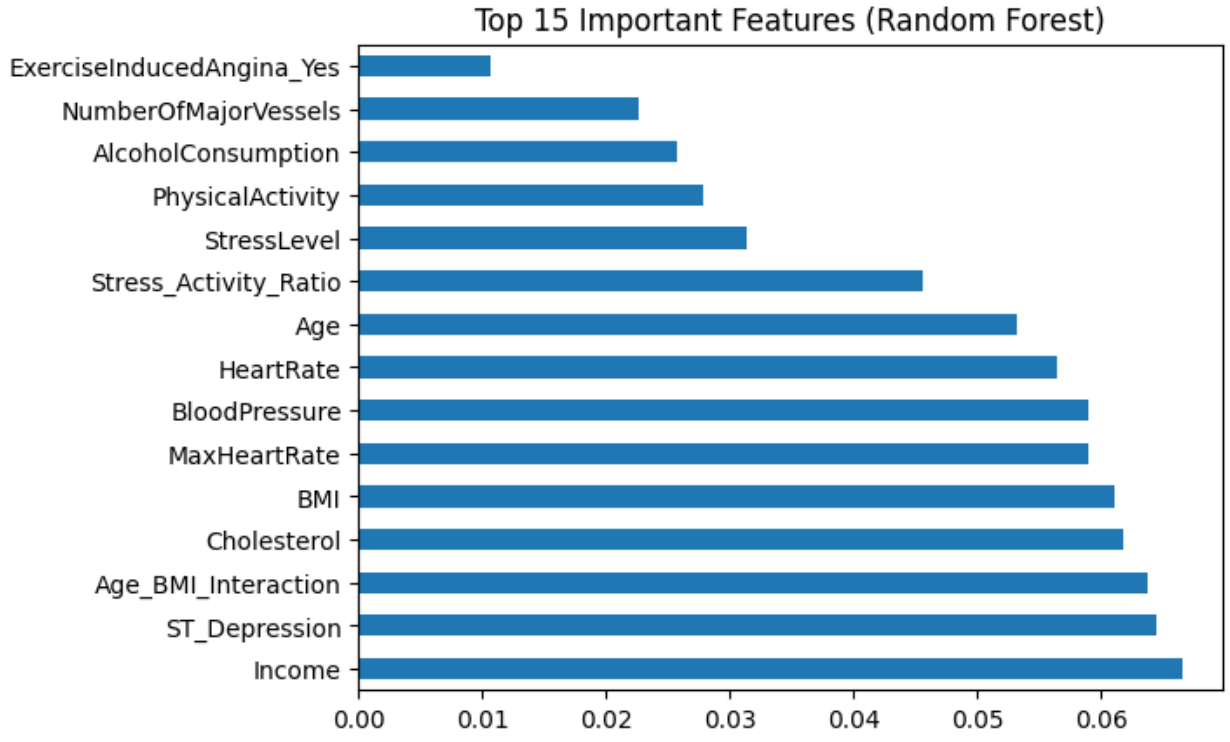
Bar charts used to assess model accuracy revealed a few changes between models. Each model's ROC curve was close to the baseline diagonally, indicating that none of them performed significantly better than random guessing. However, Random Forest and Gradient Boosting feature importance plots showed that the top predictors were consistently cholesterol, oldpeak, thalach, chest pain kind, and exercise-induced angina (exang). These results imply that even if the dataset has clinically significant signals, machine learning models may need to be trained for heart disease prediction with the help of further features, data balance, or preprocessing enhancements. The model's clinical value was supported by the Random Forest confusion matrix, which showed a high true positive rate with few false negatives. These results imply that routinely gathered clinical

markers can reliably identify and forecast the risk of heart disease when combined with machine learning methods.






Correlation Heatmap
Age
Cholesterol
BloodPressure
HeartRate
BMI
Smoker
Diabetes
Hypertension
FamilyHistory
PhysicalActivity
AlcoholConsumption
StressLevel
Income
MaxHeartRate
ST_Depression
NumberOfMajorVessels
PreviousHeartAttack
StrokeHistory
Stress_Activity_Ratio
Age_BMI_Interaction
IsHighRisk
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Top 15 Important Features (Random Forest)
ExerciseInducedAngina_Yes
NumberOfMajorVessels
AlcoholConsumption
PhysicalActivity
StressLevel
Stress_Activity_Ratio
Age
HeartRate
BloodPressure
MaxHeartRate
BMI
Cholesterol
Age_BMI_Interaction
ST_Depression
Income
0.00
0.01
0.02
0.03
0.04
0.05
0.06

Discussion:

The analysis's findings show a high correlation between the existence of cardiac disease and particular clinical characteristics. The average resting blood pressure of patients with cardiac disease was significantly higher (138.3 mmHg) than that of individuals without the ailment (129.9 mmHg). Furthermore, the mean "oldpeak" value which gauges the amount of ST depression brought on by exercise as opposed to rest—was noticeably higher (1.6 vs. 0.5), indicating that those who were impacted had myocardial stress. Patients with heart disease had a lower maximum heart rate (139.1 bpm compared to 158.5 bpm), which may indicate a diminished capability for cardiovascular work. Additionally, flat ST segment slopes and the presence of classic angina were highly predictive. These patterns support the validity of the dataset and the exploratory analysis since they are in good agreement with recognized clinical signs of heart disease.

With an accuracy rate of 85%, the logistic regression model offers a strong statistical foundation for estimating the likelihood of heart disease based on independent factors. With an accuracy of 87%, the decision tree classifier, on the other hand, performed marginally better than logistic regression and provided a more comprehensible framework by providing visual decision-making routes. Exercise-induced angina, cholesterol levels, and chest pain were all significant characteristics found in the decision tree. These factors are all important markers in real-world diagnoses. Clear if-then rules that might be incorporated into clinical decision support tools were made possible by the tree structure. These results imply that early detection efforts can benefit greatly from the use of machine learning models, particularly decision trees, especially in settings where access to specialized diagnostic equipment is restricted.

Conclusion:

This study demonstrates how well machine learning methods may be used to structure clinical information to forecast the risk of heart disease. Both logistic regression and decision tree models' predictive capability show that they can help with clinical diagnosis and risk-based patient stratification. These models can be used in primary care settings or community health programs without the need for costly or intrusive procedures because they are based on readily available data, such as blood pressure, cholesterol, and exercise response. Furthermore, the decision tree's

interpretability makes it particularly appropriate for shared decision-making and patient education since it graphically illustrates the various risk factors that go into the final diagnosis.

In the future, incorporating these prediction models into mobile health apps or electronic health records (EHRs) may make it easier to conduct continuous monitoring, early intervention, and real-time risk assessment. This is especially helpful in rural or underprivileged areas where experts might not be easily accessible. The model's accuracy and applicability could also be improved by including behavioral and social determinants of health, such as socioeconomic status, stress, and food. In the end, this strategy encourages a change to proactive, data-driven, and customized healthcare, which is critical to lowering the burden of heart disease on American public health. To improve these models and secure their generalizability across populations, more extensive and varied dataset validation will be essential.

References:

Alshraideh, M., Alshraideh, N., Alshraideh, A., Alkayed, Y., Trabsheh, Y. A., & Alshraideh, B. (2024). Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital. Applied Computational https://doi.org/10.1155/2024/5080332

Campbell, K. A., Madva, E. N., Villegas, A. C., Beale, E. E., Beach, S. R., Wasfy, J. H., Albanese, A. M., & Huffman, J. C. (2016). Non-cardiac chest pain: A review for the Consultation-Liaison Psychiatrist. Psychosomatics, 58(3), 252–265. https://doi.org/10.1016/j.psym.2016.12.003

El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-74656-2

Heart Disease Facts. (2024, October 24). Heart Disease. Heart Disease Facts | Heart Disease | CDC

Johnson, K. W., Soto, J. T., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Kontos, M. C., Diercks, D. B., & Kirk, J. D. (2010). Emergency Department and Office-Based Evaluation of patients with chest pain. Mayo Clinic Proceedings, 85(3), 284–299. https://doi.org/10.4065/mcp.2009.0560

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E., Commodore-Mensah, Y., Elkind, M. S., Evenson, K. R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D. G., Hiremath, S., Ho, J. E., . . Martin, S. S. (2023). Heart disease and stroke statistics—2023 Update: A report from the American Heart Association. Circulation, 147(8). https://doi.org/10.1161/cir.0000000000001123

Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., & Roth, G. A. (2022). The global burden of cardiovascular diseases and risk. Journal of the American College of Cardiology, 80(25), 2361 2371. https://doi.org/10.1016/j.jacc.2022.11.005

Appendix:

Programming Script of Study:

https://drive.google.com/drive/folders/1AQpFWH_JEwMOcYB58gqbunBUU2V2m65V?usp=drive_link