**Heart Attack Analysis & Prediction – Final Project**

**By Nathan Alegria, Brandon Stanyer, & Christine Kwan**

**Introduction:**

Our project aims to create models which predict if an individual is at risk of cardiovascular issues. Using patient medical data, such as blood pressure and cholesterol, we will train machine learning models to identify individuals who may be at risk of heart attack.

The dataset offers an in-depth look at factors related to heart attack risks, including patients' medical history, lifestyle, and physiological measurements. Collected from reputable medical studies and hospital records, it features indicators like age, gender, cholesterol levels, blood pressure, and smoking status. The dataset also includes labels showing whether patients had a heart attack, making it useful for analyzing patterns and developing predictive models. It is valuable for healthcare professionals, researchers, and data scientists working on cardiovascular health, risk assessment, and prevention strategies.

**About the Heart Attack Dataset:**

This dataset contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no/less chance of heart attack and 1 = more chance of heart attack

**Data Overview:**

The dataset was obtained from: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>

**Attribute Information:**

* **age:** Age of the patient.
* **sex:** Sex of the patient (0 = female, 1 = male).
* **cp:** Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic).
* **trtbps:** Resting blood pressure in mm Hg at the time of hospital admission.
* **chol:** Serum cholesterol in mg/dl.
* **fbs:** Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
* **restecg:** Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy).
* **thalachh:** Maximum heart rate achieved.
* **exng:** Exercise-induced angina (1 = yes, 0 = no).
* **oldpeak:** ST depression induced by exercise relative to rest.
* **slp:** The slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping).
* **caa:** Number of major vessels (0-3) colored by fluoroscopy.
* **thall:** Thalassemia (0 = normal, 1 = fixed defect, 2 = reversible defect).
* **output:** Target variable (0 = less chance of heart attack, 1 = more chance of heart attack).

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

**Data Preprocessing & Cleaning:**

In this stage, the dataset was already well-structured and free of missing values, so minimal data cleaning was required.

A screenshot of a computer program

Description automatically generated

**Data Exploration:**

In the initial data exploration phase, we visualized the distribution of each attribute to understand the dataset better. By generating histograms for numerical variables and bar plots for categorical variables, can examine the spread, central tendencies, and potential outliers in the data. These visualizations provided valuable insights into the underlying patterns and relationships, helping us identify any skewness or irregularities that could influence the subsequent analysis.

A graph of age and age

Description automatically generated

The graphs show the distribution of **Age** and **Sex** in the dataset:

* **Age**: The distribution is mostly normal, with most individuals between 50 and 60 years old. There are fewer individuals under 40 and over 70, indicating the data is concentrated in middle-aged groups.
* **Sex**: The distribution is imbalanced, with more males (labeled "1") than females (labeled "0"), showing a male-heavy dataset.

A comparison of different types of pain

Description automatically generated

* **Distribution of Chest Pain Type:** The majority of individuals experience chest pain type 0 (typical angina), followed by type 2 (non-anginal pain). Fewer individuals are categorized with chest pain types 1 (atypical angina) and 3 (asymptomatic). This suggests that typical angina is the most common chest pain type in the dataset.
* **Distribution of Resting BP (Blood Pressure):** The resting blood pressure distribution is roughly normal, with most values ranging between 120 and 140 mmHg. There are fewer cases at the extremes, with very few individuals having resting BP below 100 or above 160 mmHg. This indicates that the data is concentrated around typical resting BP levels.

A comparison of a graph

Description automatically generated with medium confidence

* **Distribution of Cholesterol:** The cholesterol levels are approximately normally distributed, with most values falling between 200 and 300 mg/dL. There are fewer individuals with very high cholesterol levels above 400 mg/dL, indicating that most patients in the dataset have moderate cholesterol levels.
* **Distribution of Fasting BS (Blood Sugar):** The distribution is heavily skewed, with the majority of individuals having fasting blood sugar levels below 120 mg/dL (labeled as 0). Only a small portion of individuals have elevated fasting blood sugar levels above 120 mg/dL (labeled as 1), indicating that high blood sugar is relatively uncommon in this dataset.

A comparison of a graph

Description automatically generated

* **Distribution of Resting ECG:** The resting ECG results show a clear division into three categories. Most individuals fall into categories 0 (normal) and 1 (ST-T wave abnormality), while only a few individuals are in category 2 (left ventricular hypertrophy). This indicates that most patients have either normal ECG results or minor abnormalities.
* **Distribution of Max HR (Maximum Heart Rate):** The maximum heart rate achieved by individuals follows a normal distribution, with most values between 140 and 160 beats per minute.

A graph of a normal and a normal and a normal

Description automatically generated with medium confidence

* **Distribution of Exercise IA (Exercise-Induced Angina)**: The majority of individuals (labeled as "0") do not experience exercise-induced angina, while a smaller group (labeled as "1") does. This indicates that most patients in the dataset do not suffer from angina during exercise.
* **Distribution of ST Depression:** The ST depression values are skewed to the right, with most individuals having low ST depression levels near 0. A smaller number of individuals have higher ST depression values, which suggests that significant ST depression during exercise is less common in the dataset.

A comparison of a number of objects

Description automatically generated with medium confidence

* **Distribution of SLP (Slope of the Peak Exercise ST Segment):** The majority of individuals fall into categories 1 and 2, indicating a flat or downsloping ST segment. Fewer individuals are in category 0 (upsloping), suggesting that most patients exhibit either flat or downsloping ST segments during exercise.
* **Distribution of the Number of Major Vessels:** Most individuals have 0 major vessels colored by fluoroscopy, with decreasing frequencies as the number of vessels increases. Very few individuals have 3 or 4 major vessels affected, indicating that significant blockage in multiple vessels is less common in the dataset.

A graph of distribution of thalassemia

Description automatically generated

* **Distribution of Thalassemia:** The majority of individuals fall into category 2 (reversible defect), followed by category 3 (fixed defect). Very few individuals are in category 0 (normal), indicating that most patients in the dataset exhibit some form of thalassemia, with reversible and fixed defects being the most common types.

**Logistic Regression Model**

Logistic regression is a method used to predict whether someone is at risk of a heart attack based on different health factors, like age, cholesterol, and chest pain type. Since the outcome is either "at risk" or "not at risk," logistic regression is a good fit for this kind of problem. The main goal is to use the model to make predictions and understand which factors are most important in determining heart attack risk.

**A screenshot of a computer

Description automatically generated**

**A blue squares with white text

Description automatically generated**

**Analysis of Accuracy:** The model's performance, reflected in the confusion matrix, demonstrates good results for both true positives (42) and true negatives (32), which indicate correct predictions for individuals with and without heart attack risk. It is a good sign they outnumber the false positives (9) and false negatives (8).

The overall accuracy of 81% means that the model correctly classified 81% of the cases, combining both true positives and true negatives.

**ROC Curve**

A graph of a receiver operating characteristic curve

Description automatically generated

**Analysis of ROC Curve:** The AUC (Area Under the Curve) is 0.88, which indicates that the model performs very well. An AUC of 0.88 means the model has an 88% chance of correctly distinguishing between a person who is at risk of a heart attack and a person who is not. This suggests that the model is quite effective at making predictions.

**Analysis & Observation of Logistic Regression Heatmap:**

A diagram of heatmap

Description automatically generated

The heatmap helps to visualize which attributes are more closely related to heart attack risk and how other features are interrelated. It gives a clear sense of which factors might be more important for predicting heart attack risk.

* Blue areas show positive relationships (darker blue means stronger positive correlation).
* Red areas show negative relationships (darker red means stronger negative correlation).
* Lighter colors indicate weaker or no relationship.

**Heart Attack Risk:**

* Number of Major Vessels and Exercise IA have the strongest positive relationships with heart attack risk. This means people with more affected vessels or who experience exercise-induced angina are more likely to have a heart attack.
* ST Depression and Chest Pain Type also have moderate positive correlations with heart attack risk.
* Max HR (maximum heart rate) has a moderate negative correlation, meaning lower max heart rates are linked to higher heart attack risk.

**Strong Positive Correlations:**

* **Number of Major Vessels and Heart Attack Risk:** A higher number of affected vessels correlates with increased heart attack risk.
* **Chest Pain Type and Heart Attack Risk:** Certain types of chest pain are linked to higher heart attack risk.
* **ST Depression and Heart Attack Risk:** Increased ST Depression correlates with higher heart attack risk.
* **Exercise IA and Heart Attack Risk:** Exercise-induced angina is strongly associated with increased heart attack risk.

**Negative Correlations:**

* **Max HR and Heart Attack Risk:** Higher maximum heart rate is linked to lower heart attack risk.
* **Exercise IA and Max HR:** Individuals with exercise-induced angina tend to have lower maximum heart rates.
* **Age and Max HR:** Older individuals tend to have lower maximum heart rates.

**KNN (K-Nearest Neighbor) Model**

A screenshot of a computer screen

Description automatically generated

A blue squares with white text

Description automatically generated

**Analysis of Accuracy:**

The overall accuracy of 88% means that the model correctly classified 88% of the cases, combining both true positives and true negatives.

The KNN model shows strong performance based on the confusion matrix, with 36 true positives (correctly identified heart attack risks) and 31 true negatives (correctly identified individuals without heart attack risk). These correct predictions outnumber the false positives (4) and false negatives (5), indicating that the model is accurate in distinguishing between individuals at risk and those not at risk.

Compared to the Logistic Regression model, the KNN model appears to be more balanced, as it has fewer false positives and false negatives, suggesting a more even distribution of correct classifications.

A graph with a line

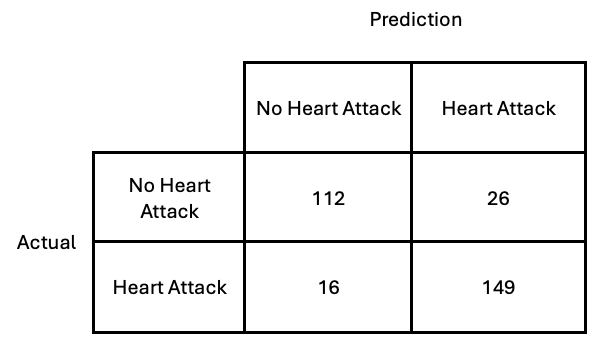
Description automatically generated

**Analysis of ROC Curve:** The ROC curve for the KNN model suggests that it is effective at predicting heart attack risk, with a high AUC score of 0.89. This means that the model is able to correctly classify heart attack risk in the majority of cases, striking a good balance between true positives and false positives.

**Support Vector Machines Model**

The support vector machines (SVM) model proved to be a very accurate model for predicting heart attacks. The model shows minimal bias in its predictions. However, the challenge in interpreting the SVM model may not make it suitable for this application. SVM, especially within spark, does not allow for many of the helpful visualizations shown with the other models.

**Model Metrics:**



Test Accuracy: 0.861

Precision: 0.862

Recall: 0.861

F1-Score: 0.860

**Analysis of Model:**

The SVM model achieved an accuracy of approximately 86%. Given the context of heart attack prediction and the intricacies of human health, this is quite an effective model. The model made mistakes on 16 of the 165 patients who had a heart attack. Given that, the model was also quite unbiased in its errors, with 26 false positives.

Overall, this LinearSVC model has good performance on this task.

**Decision Tree (Bagging) & Random Forest**

Bagging trains multiple decision trees on random subsets of data and uses all features for splits, combining their predictions to reduce overfitting. Random Forest extends this by selecting a random subset of features at each split, which reduces tree correlation and improves accuracy. Random Forest is generally more robust, handles overfitting better, and provides feature importance insights, making it more effective for complex tasks.

However, both Decision Tree (Bagging) and Random Forest perform comparably to other algorithms, with accuracy ranging between 74% and 76%. Random Forest slightly outperforms Decision Tree (Bagging).

For the Random Forest Confusion Matrix:

* True Positive = 30
* True Negative = 39
* False Positive = 11
* False Negative = 11

A screenshot of a graph

Description automatically generated

For the Decision Tree (Bagging) Confusion Matrix:

* True Positive = 31
* True Negative = 37
* False Positive = 10
* False Negative = 13

A graph with blue squares

Description automatically generated

One of the benefits of using Random Forest is the ability to understand feature importance. This metric shows how much each feature contributes to a model's predictions. In the models, feature importance is determined by how often a feature is used to split data or by how much it reduces error. This helps identify the most influential features, making the model more interpretable and enabling dimensionality reduction by removing less important variables. See the Random Forest Tree below for more further understanding of the result for the feature importance.

A diagram of a structure

Description automatically generated

Features such as Thalassemia, Resting BP, and Cholesterol appear early in the tree, suggesting they are significant contributors to the model’s decision-making process. In terms of performance, the model achieves squared errors close to 0 in many leaf nodes, indicating that predictions in those nodes are highly accurate for the samples falling into them. Finally, both the Random Forest and decision tree models achieved an accuracy of more than 70%, which indicates that the models are performing reasonably well in predicting outcomes based on the data.

**Conclusion**

This project tested 4 different types of machine learning models for predicting heart attacks. Each of the models did quite well with only 14 health measures provided. Maximum heart rate and chest pain type appear to be the most relevant factors when assessing a patient's risk of a heart attack, based on the correlation assessment. With all 14 variables used, the machine learning models were able to achieve accuracy over 80%. Given the complexity of the human body, this seems to constitute an effective model.

The most effective models for this task were K-Nearest Neighbor (KNN) and Support Vector Machines (LinearSVC). KNN seems the best suited for this task. Not only did it achieve the highest accuracy of the models tested, it also can be interpreted well. With Principal Component Analysis (PCA), we can easily understand which variables most affect the variance in heart attacks. In addition, PCA allows us to see which factors lead to patients being grouped together in clusters.

Overall, the prediction of heart attacks is suitable for machine learning algorithms. This project proposes multiple types of models which achieve effective accuracy measures, KNN being the preferred. To improve this project, it would be helpful to tune the parameters of the KNN model and carefully analyze the model using PCA to learn more. Since this dataset is relatively small, it would also be helpful to gather more data and see if the models perform similarly.