# Framingham Dataset

**Exploratory Data Analysis (EDA)**

"Framingham" heart disease dataset includes over 4,240 records,16 columns and 15 attributes. The goal of the dataset is to predict whether the patient has 10-year risk of future (CHD) coronary heart disease

**Heart Failure Attributes**

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

**Dataset Description**

The dataset under investigation contains information on individuals, including features such as age, education level, gender, smoking status, and various health-related measurements (e.g., cholesterol levels, blood pressure, BMI, and glucose levels). The dataset also includes a crucial target variable, "TenYearCHD," which indicates whether an individual is at risk of developing CHD within the next ten years (1 for at risk, 0 for not at risk).

**Objectives**

The primary objectives of this project are as follows:

1. Data Exploration: Conduct a thorough exploration of the dataset to understand its structure, identify missing values, and gain insights into the distribution of key variables.
2. Feature Analysis: Analyze the dataset's features to determine their suitability for predicting Ten-Year CHD risk. Categorize features as numeric, categorical, ordinal, or nominal.
3. Dependency Analysis: Employ statistical tests, including the Chi-squared test for categorical features and t-tests or ANOVA for numeric features, to assess the dependency of each feature on the Ten-Year CHD target variable.
4. Data Preprocessing: Prepare the dataset for modeling by handling missing values, encoding categorical variables, and scaling or normalizing numeric features.
5. Model Building: Develop predictive models using machine learning algorithms to predict Ten-Year CHD risk based on the selected features.
6. Model Evaluation: Assess the performance of the developed models using appropriate evaluation metrics (e.g., accuracy, precision, recall, ROC curve) to determine their predictive capability.
7. Interpretation: Interpret the results to identify the most influential factors contributing to Ten-Year CHD risk.

A white paper with black text

Description automatically generated

# UCI Dataset

This database 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 date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

**Column Descriptions:**

1. Id: (Unique id for each patient)
2. age: (Age of the patient in years)
3. origin: (place of study)
4. sex: (Male/Female)
5. cp: chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
6. trestbps: resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
7. chol: (serum cholesterol in mg/dl)
8. fbs: (if fasting blood sugar > 120 mg/dl)
9. restecg: (resting electrocardiographic results)- Values: [normal, stt abnormality, lv hypertrophy]
10. thalach: maximum heart rate achieved
11. exang: exercise-induced angina (True/ False)
12. oldpeak: ST depression induced by exercise relative to rest
13. slope: the slope of the peak exercise ST segment
14. ca: number of major vessels (0-3) colored by fluoroscopy
15. thal: [normal; fixed defect; reversible defect]
16. num: the predicted attribute (target)

A graph with blue lines and numbers

Description automatically generatedA graph with blue and white lines

Description automatically generated

## Example Analysis

Best\*<https://www.kaggle.com/code/rushikeshmuley/coronary-heart-disease-prediction-ipynb>

<https://www.kaggle.com/code/xleepy/heart-csv-preprocessing-classification>

<https://www.kaggle.com/code/dhanishahahaha/feature-engineering-handling-missing-data>

MIT UCI

<https://www.kaggle.com/datasets/krishujeniya/heart-diseae/data>

**Other**

[1,888 records merged form 5 publicly available datasets - 14 features](https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset)

# Brandon Simmons Paper

**Brandon Simmons paper**

For a number of unknown reasons to this researcher, the original raw source data became corrupted and was never re-uploaded to the repository.

Using Random Forest and logistic regression, we determine the best dataset for machine learning and statistical analysis: the Cleveland data on a reduced set of six features.

Three of which are statistically significant at explaining or classifying patients as "Heart Disease". They are thalach (maximum heart rate), oldpeak, and cp (chest pain).

Our work shows how to fix the Kaggle data as it becomes essentially the Cleveland data. We determine if there is any relation between the Statlog data and the Cleveland data.

Then, we identify both of these data from the most recently published (combined) data on heart disease.

Finally, we analyze these data using logistic regression along with random forest feature selections to determine the best models with the least AI criterion. We conclude that, the Cleveland data on a reduced (features) model is the best statistical model.

Approach

1. although the data is supposed to have originated from the University of California Irvine Machine Learning Repository archive, the predictor’s encoding is backward. This led to misleading interpretations and incorrect conclusions about the features relation with the target variable. However, this data is very similar to Cleveland data which is also found in the UCI repository. Except, the Cleveland data has 6 missing values and the target variable has 5 levels. With the right encoding, we conclude that Kaggle (with target 1=No disease; 0=Disease) is the Cleveland. Still, we found another complete dataset in another archive of the UCI repository, called Stalog which is similar to Cleveland except in size.
2. Compare the Cleveland data (303 obs.) to the Statlog data (270 obs.). We use “Pandas Profiling” to access overall data content such as size and variable names, types and their distributions. Then, we compare pairwise, the distribution of each feature, and each feature against the target variable. We see very similar associations between each feature and the target and upon further investigations, we found that the Statlog data is a proper subset of the Cleveland data.
3. Find the remaining three datasets that were corrupted in the UCI dataset repository. To do so, we needed to first identify the Cleveland (and thus the Statlog datasets) from the most recently published data, ORIG—We merge the Cleveland data with ORIG data and remove all duplicated records. In which case, any record that appears in both Cleveland, and ORIG is removed. The remaining dataset is called MISS and perhaps represents the (3) missing datasets that were deemed corrupt in the UCI repository. Unfortunately, we are unable to authenticate the origin of the ORIG data. Despite our attempt to reach out to the originator, we received no response. So, we proceed to further investigate the content of the MISS data. We found that about 25% (303 observations) are duplicated records.
4. Investigate the MISS data by exploring the distribution of its features, and the features associations with the target variable. We found that the data has skewed significantly the gender, and unusually high records of ’0’ values for cholesterol level.
5. Analyze each of the three datasets: CLEVELAND, MISS, and STATLOG. We build a logistic regression on each of the data, and using Random Forest algorithm we select the most important features from each data against the target. The process led to building some reduced models and determine the best model for statistical analysis.
6. Criteria 1: Blood Flow

* CAD is defined as the obstruction of the blood flow to the heart muscles through the coronary arteries by plaque accumulation on the walls which indefinitely leads to heart attacks[12]. The CAD mortality rate visually presents the stats for 35 and older for both sexes and all ethnicity [3]. North Carolina falls in the range 148.6 - 167.9 with the estimated value of 165.2.

Criteria 2: Structure - Heart Valve Disease (HVD)

* Heart Valve Disease (HVD) involves the issues occurring with the four values that renew and direct blood flow for a smooth transition within the circulatory system[12]. In essence, when one or more of the valves is improperly functions then the patient’s condition is classified as HVD. Symptoms include stenosis, which is the limiting of muscle mobility due to valves fusing; Valve Leakage which causes blood to flow back into the heart; and lastly Atresia, which absence of valve opening

Criteria 3: Overall Function - Arrhythmias & Heart Failure

**Contributing Factors**

main criteria that is represent in the data sets are blood flow. It is apparent these following factors contribute to the number of heart disease cases within the United States in general. The factors of age, sex, and geographic location increase the risk of a heart attack or stroke[19]. It is also interesting that our heart age may be higher than our actual age due to the influences of diet, stress, activity and heritage.

Until the age of 45 years old and older, medical personnel prescribe the public to have a routine heart health check at least every two years. The condition of participants within the data sets pertains to the data retrieved from testing. At the time of data collection traditional risk factors for CAD are cholesterol levels, blood pressure, blood sugar, and pain developed from exercise recorded during the thallium test.

A table of medical information

Description automatically generated

Figure 1: 12 Selected features for Simmons

A close-up of a blood sugar level

Description automatically generatedA close-up of a medical report

Description automatically generated

A screenshot of a table

Description automatically generated

A white text with black text

Description automatically generated"Thallium Stress Test" also known as "Nuclear Stress Test" or "Cardiac Test" benefits heart disease research by analyzing the condition of blood flow [14]. The gamma camera through nuclear imaging tracks the participant’s blood flow which carries a sample amount of Thallium, radioactive isotope

A close up of a text

Description automatically generated

## Cleveland vs Kaggle Data (Simmons)

The original Cleveland data is called processed.cleveland.data, which can be downloaded from the UCI Machine Learning repository site (in “Data Folder”) and exported as a txt file. It was donated in July 1988 and currently has 1,471,444 web hits. This data has no header, and yet the headers information were provided on the website (under Attribute Information)—There are 303 records and 14 features (5 numerical, 9 categorical) information were provided which helped determine each column header in the order listed in the Table B, below. Although the Cleveland data features share similar distributions to those of Kaggle (see Appendix I), there are a couple of differences that we list below: (a) Contrary to the Kaggle data, the original Cleveland data has 6 missing values (recorded as ‘?’): 4 missing values are found for the number of blood vessels (ca) feature and 2 missing values for the thalassemia (thal) feature. Because each feature is categorical, it appears that, each of their missing values was replaced in the Kaggle data with a closest value generated by the Nearest Neighbor Algorithm. However, since these two features do not appear in a larger dataset (ORIG) which we explore later, for practical comparative analysis, decided to drop them.

# Background

Mayo Clinic organization have further defined heart disease as: “generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart’s muscle, valves or rhythm, also are considered forms of heart disease.