INSTITUTE OF DATA
WK 7

MINI PROJECT 3

HEART DISEASE DATASET (UCI)

- The evaluated dataset contains information regarding the presence of heart disease in patients across 4 databases - Cleveland, Hungary, Switzerland and the VA Long Beach. The original data contained 76 attributes (columns), however, all the published experiments focused on a subset of only 14 attributes. The Cleveland database was refined to distinguish the presence from absence of heart disease, rather than the degree of presence (which was evident in the three other unrefined databases). This is the set of data that will be analysed within this project as it signifies the potential for binary classification.
- Dataset characteristics: Multivariate, structured. 303 patients, 14 refined attributes.
- The database was found from the UCI Machine Learning Repository, but effectively retrieved through Kaggle's Public API.
- Original authors/ principal investigators responsible for the collection of the dataset are as denoted:

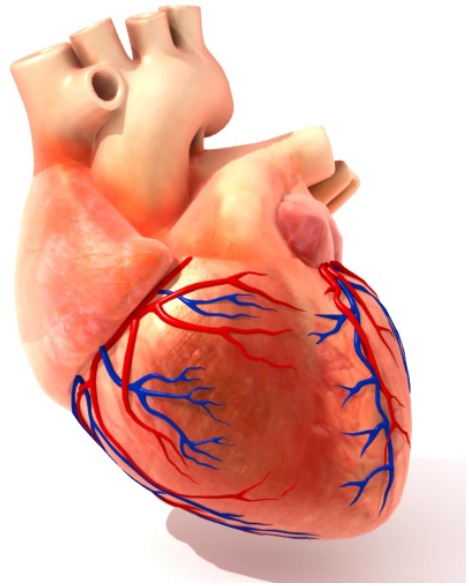
  Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

  University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

  University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

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

  Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779 (Date donated: 01/07/1988)
- PROBLEM:
  - Can the presence of heart disease be predicted using machine learning models?
  - How accurate will these models be?



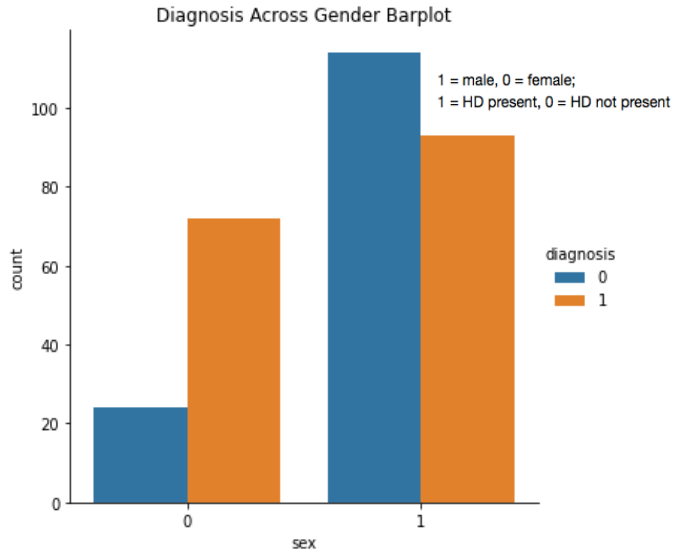
DATASET ATTRIBUTES (FOR REFERENCE)

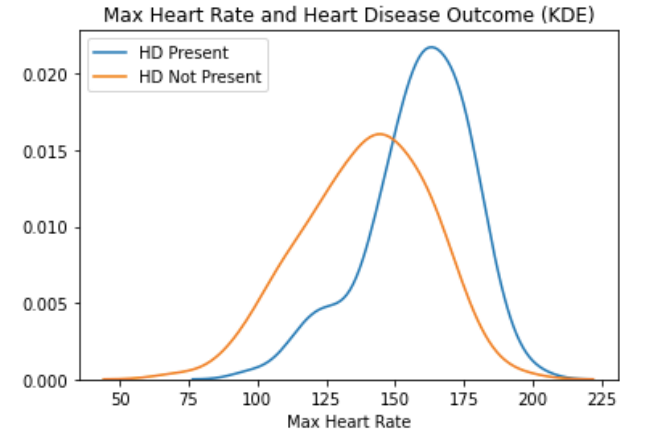
Attribute	Details
Age	age in years
Sex	gender (1 = male, 0 = female)
Chest Pain	chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
Resting Blood Pressure	resting blood pressure (in mm Hg on admission to the hospital)
Cholesterol	serum cholestoral in mg/dl
Fasting Blood Sugar	when fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
Rest ECG	resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
Max Heart Rate	maximum heart rate achieved
Exercise Angina	exercise induced angina (1 = yes, 0 = no)
ST Depression	ST depression induced by exercise relative to rest
ST Slope	the slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
Major Vessels	number of major vessels (0-3) colored by flourosopy
Thalassemia	thalassemia, a blood disorder (3 = normal; 6 = fixed defect; 7 = reversable defect)
Diagnosis	diagnosis of heart disease (0 'No' = < 50% diameter narrowing, 1 'Yes' = > 50% diameter narrowing)

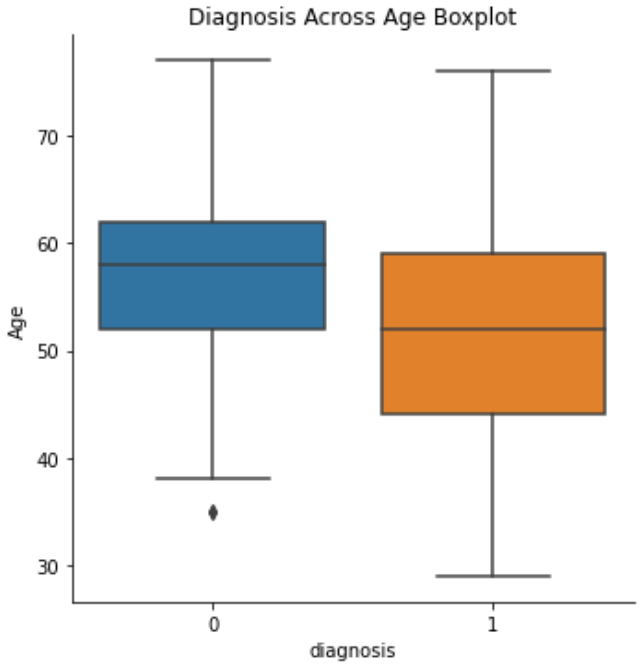
DATA PRE-PROCESSING AND ANALYSIS

- SIGNIFICANT FINDINGS:

1. The proportion of females that had heart disease (HD) present was higher than that of males. Out of 96 females, 72 have HD (75%), out of 207 males, 93 have HD (45%). Note: the dataset was skewed towards male samples.
2. The KDE (kernel density estimate) visualizes the distribution of maximum heart rate across patients, distinguished by diagnosis. It is evident that those with HD present generally experience a higher heart rate than patients without HD.
3. The age of patients within this dataset ranged from 29 to 77. Interestingly, the average age of those with HD present was lower than that of those without HD. However the distribution of age across patients with HD present was far greater.







DATA STORAGE

- The Heart Disease (UCI) sourced through Kaggle's public API was stored locally within a SQLite database.
  - The retrieved data was structured in a relational tabular format (relationships present between incidents and attributes).
  - SQLite is a relational database management system. In other words, it is a program that enables a user to create, update and administer relational datasets.
  - SQLite functions with these structured datasets in assisting the ability to store, process and access each discrete field.

MACHINE LEARNING MODEL

FEATURE ENGINEERING

- Output (target variable):
  - Diagnostic result, of the patient regarding heart disease
  - This output is binary, distinguishing presence (1) from absence (0) of the condition.
- Input (predictor variables):
  - As the dataset has already seen significant feature selection (76 to 14 attributes) and processing I am reluctant to cull anymore. For the sake of the project, I will use features with a correlation to diagnosis greater than +/- 0.2
  - All predictor features remain numerical, so no dummy variables or encoded mapping is necessary
  - The resultant feature vector is as follows (head):

	chest_pain	max_heart_rate	st_slope	age	sex	thalassemia	major_vessels	st_depression	exercise_angina
0	3	150	0	63	1	1	0	2.3	0
1	2	187	0	37	1	2	0	3.5	0
2	1	172	2	41	0	2	0	1.4	0
3	1	178	2	56	1	2	0	0.8	0
4	0	163	2	57	0	2	0	0.6	1

LOGISTIC REGRESSION:

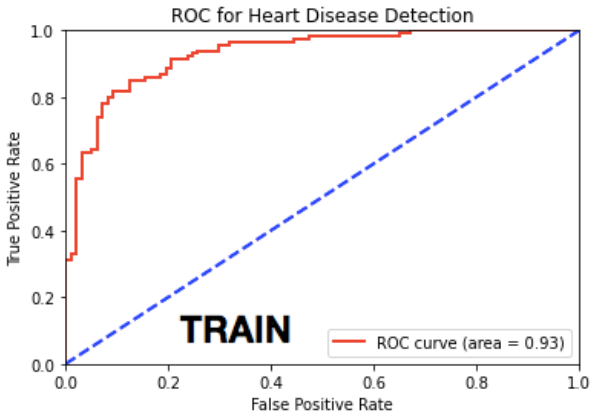
- As the target variable of interest is binary and categorical in nature, linear regression would not be a feasible model as predictions can fall external to the discrete output (1 or 0)
- As a model, logistic regression adapts so that predicted y (target variable) remains discrete.
- Logit function: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \boldsymbol{X}$
- Calculate predicted y (y hat): $\hat{y} = \beta_0 + \beta_1 \boldsymbol{X}$
- Calculate p (probability of True, where 0 < p > 1):

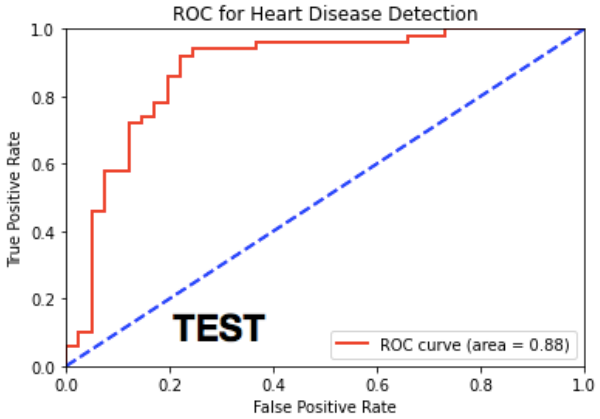
$$p = \frac{e^{\hat{y}}}{e^{\hat{y}} + 1} = \frac{1}{1 + e^{-\hat{y}}}$$

EVALUATION OF CLASSIFICATION RESULTS

TRAIN AND TESTING RESULTS:

- Train Accuracy Score: 85.85%
- Test Accuracy Score: 82.42%
  - Train and test set both predicted as well as each other, hence overfitting seems to be a low possibility.
- Receiver Operating Characteristics (ROC) curve:
  - Comparison of True Positive and False Positive Rates





Confusion Matrix:

- Depicts the counts of the actual values within the data and the predicted values drawn from the logistic regression model

	Predicted HD	Predicted Healthy
HD Present	33	8
HD Not Present	8	42

Classification Report and Evaluation Scores:

Evaluation Score	Test Set Result
Classification Error (Misclassification Rate)	17.58%
Recall (Sensitivity)	84%
Precision	84%
Specificity	80.49%

	precision	recall	f1-score	support
0	0.80	0.80	0.80	41
1	0.84	0.84	0.84	50
accuracy			0.82	91
macro avg	0.82	0.82	0.82	91
weighted avg	0.82	0.82	0.82	91

CONCLUSION:

- Based on the above modelling, it can be determined that Logistic Regression can be used to predict the presence of heart disease amongst patients in this scenario. This is validated by the strong classification accuracy score of 82% (test set), and coherently low misclassification rate of 18%.
  The Recall and precision of the model also both sit at 84%, further validating the dependency of the model as it suggests that the model is very sensitive in accurately predicting positive instances.
- Based on similar patient attributes within a medical database, this classification ML model may predict health risks of this nature into the future.