Machine Learning  
Final Report

By: Mohammad Ghosheh & Wisam Yaghmour

**1.1 Introduction**

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.  
  
There are two types of machine learning that we learnt in class, supervised and unsupervised learning, both have their uses, the former depends on the real outcome of the tested data, the latter computes the output using solely the input data, The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.  
  
In this project we will use adaboost and logistic regression and compare the results of both of them, as per their definition, they are defined as follows:

* **Logistic Regression:** it is a statistical model used in machine learning for binary classification tasks, which are tasks with two possible outcomes. It's a type of regression analysis but is suited to models where the dependent variable is binary or dichotomous, meaning it can take on one of two possible states, such as 0/1, Yes/No, or True/False.
* **Adaboost:** short for Adaptive Boosting, is a powerful ensemble machine learning algorithm that constructs a strong classifier by combining multiple weak classifiers. It begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but adjusts the weights of incorrectly classified instances such that subsequent classifiers focus more on difficult cases. This process continues until a predefined number of weak learners have been created or no further improvement can be made on the training dataset. The final model is a weighted sum of the weak classifiers, forming a strong, robust model. We used weak learners and decision trees for Adaboost.

**2.1 Dataset**

For the used dataset, We used a Dataset called “Hearts” which has a binary classification and according to the features in this dataset, it tells if the person with the supplied data will experience a heart stroke or no.

The Dataset consists of 13 features and according to training the model on them and their respected outcomes, the model trains on a specific part of the data, and then tests on the minor part of the data.  
the features are: age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall.

age: The person's age in years.

sex: The person's sex (1 = male, 0 = female).

cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic).

trtbps: The person's resting blood pressure (in mm Hg on admission to the hospital).

chol: The person's cholesterol measurement in mg/dl.

fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false).

restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria).

thalachh: The person's maximum heart rate achieved.

exng: Exercise induced angina (1 = yes; 0 = no).

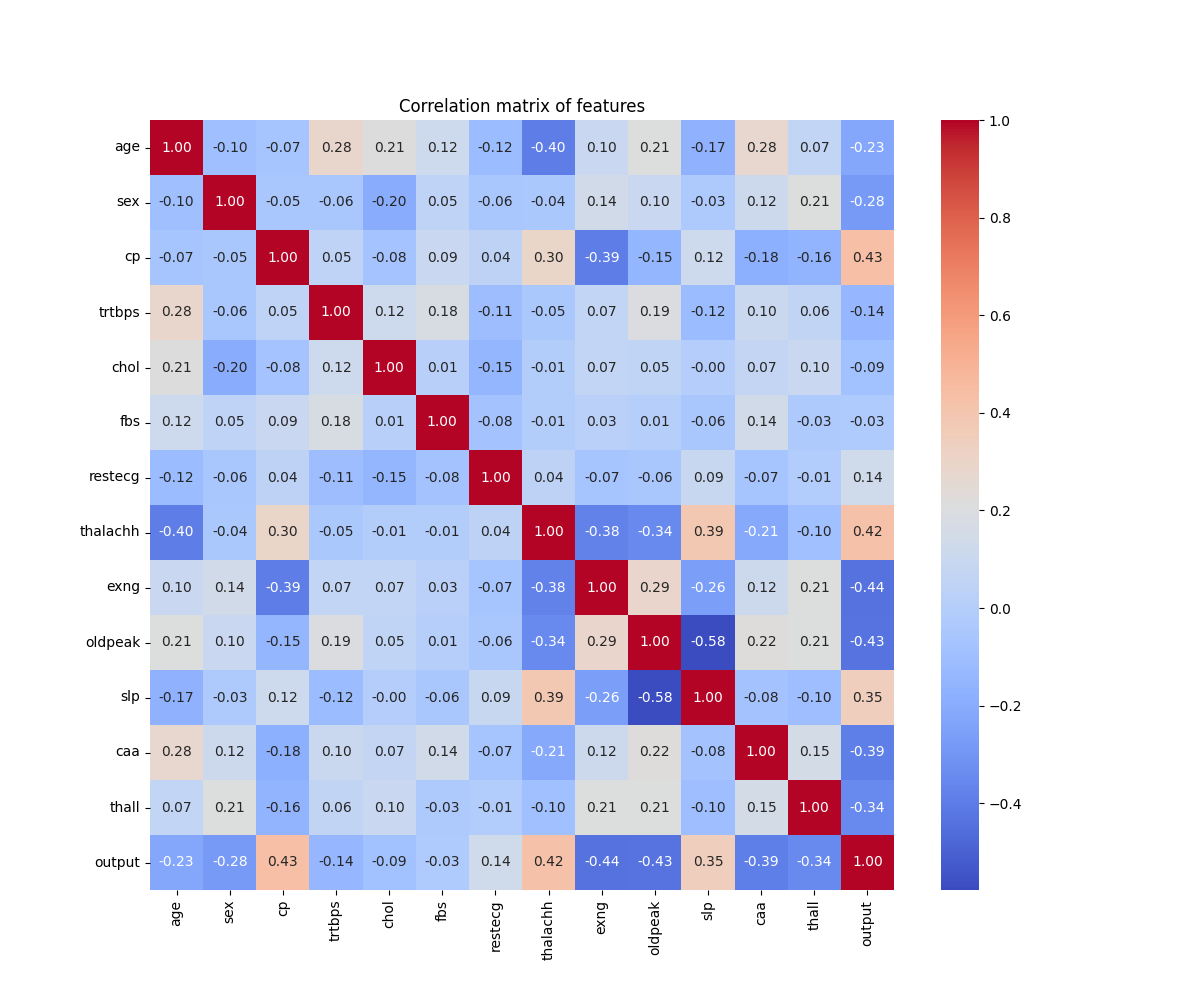
oldpeak: ST depression induced by exercise relative to rest.

slp: The slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping).

caa: Number of major vessels (0-3).

thall: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect).

Here is the correlation matrix for the dataset.



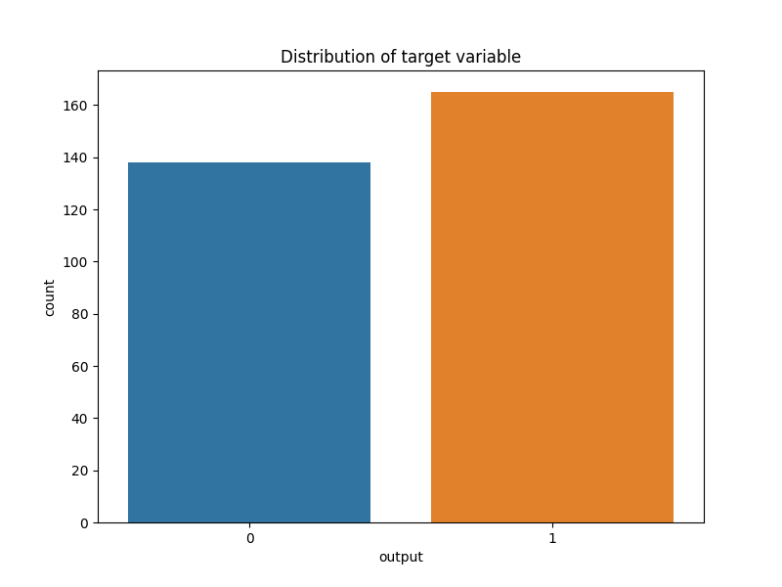
If the relation between two features is close to 1, then the relationship is positive relation, if it is close to -1 then the relation is negative, else if close to 0 then there is no relationship between the two features.

Correlation does not imply causation, this means that if two features have the relation of 1 for example, it does not really say that one feature caused the other.

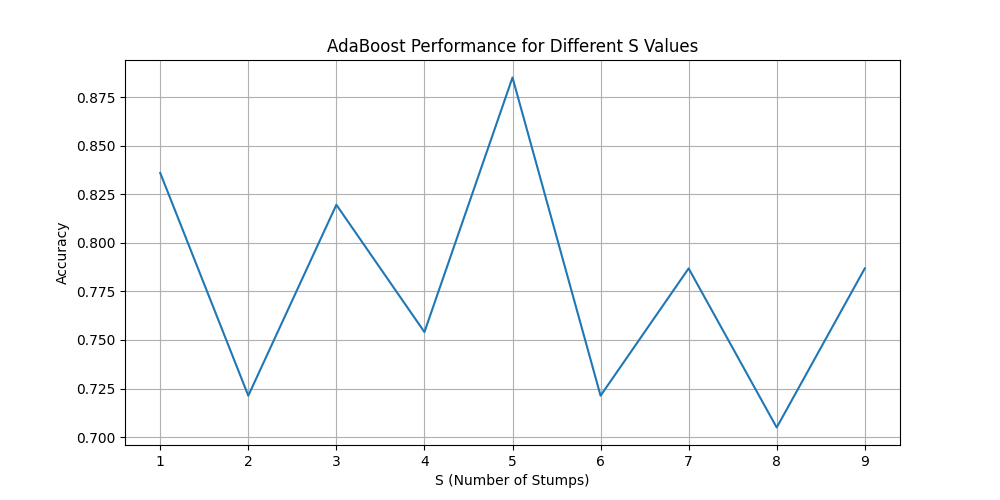
**3.1 Data Distribution**

We split the data to training and testing, in our case, 80% of the data is for training and 20% is for testing.

This is the distribution of the target variable:



**4.1 Performance**

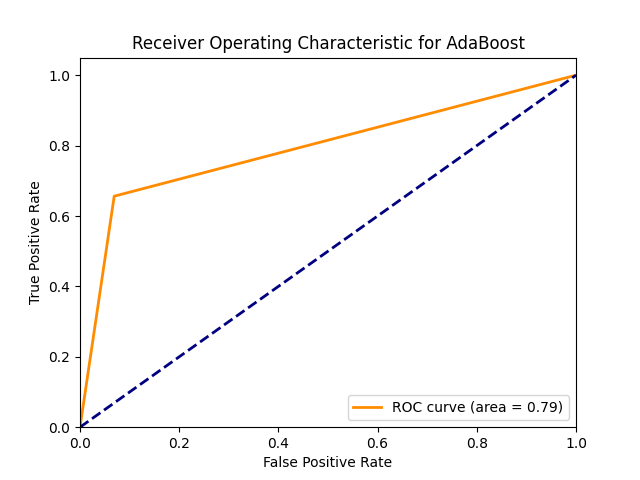
After running both algorithms on the test data which is 20% of the dataset, the results we got were interesting, Logistic regression gave 88.5% accuracy on 1000 iterations, but Adaboost on the other hand, was not consistent with it’s accuracy and it depended on the number of weak learners and tree depth, here is the graph of how the number of weak learners affected Adaboost’s accuracy:

As we can see, using 5 weak learners the accuracy was best, the accuracy is low before because of underfitting which is implementing not enough weak learners for this dataset, and also the accuracy is low after 5 because of overfitting which gets the noise in the data and uses complex running algorithm for data that does not need all that, therefore releasing unreal outcomes.

Now attached is the roc curve for both algorithms, The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model. It is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

The ROC curve helps us understand the trade-off between sensitivity (or TPR) and specificity (1 - FPR):

A graph of a logistic curve

Description automatically generated

**5.1 Results**

The quality of both algorithms are not decided on just accuracy, there are cases where the accuracy is high just because the true positives or the false positives are high without the regard of the other variable, we examine the data using the conversion matrix for each algorithm which shows the numbers of the true positives, true negatives, false positives and false negatives:

A blue squares with white text

Description automatically generatedA blue squares with white text

Description automatically generated

|  |
| --- |
|  |