**Background of the problem**

* The accurate diagnosis of heart disease can reduce the risk of serious health problems, whereas an inaccurate diagnosis can be fatal. The Kaggle Heart Attack Analyses & Prediction Dataset outcomes and analysis are compared in this study using a variety of machine learning algorithms. The dataset has 14 key attributes that will be used in the research. The accuracy and confusion matrix is used to achieve and validate a number of promising results. With the aid of machine learning, 80.3% accuracy was attained.

**Motivation for solving the problem (how important it is, etc.)**

* The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

**Solution methodology and evaluation metrics.**

**Solution methodology:**

* Are we checking on missing values and how are we going to proceed further?
* Checking for skewness and normal distribution of data?
* Considering Decision tree and logistic regression for the dataset – Proposing.

**Evaluation metrics:** Confusion matrix and accuracy.

**Description of your dataset**

About this dataset

Age : Age of the patient

Sex : Sex of the patient

exang: exercise induced angina (1 = yes; 0 = no)

ca: number of major vessels (0-3)

cp : Chest Pain type chest pain type

Value 1: typical angina

Value 2: atypical angina

Value 3: non-anginal pain

Value 4: asymptomatic

trtbps : resting blood pressure (in mm Hg)

chol : cholestoral in mg/dl fetched via BMI sensor

fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

rest\_ecg : resting electrocardiographic results

Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach : maximum heart rate achieved

target : 0= less chance of heart attack 1= more chance of heart attack

**Comparison of at least two algorithms; for each one, conduct at least two experiments, varying one or more parameters (e.g., number of trees, minimum number of data points in a leaf; number of hidden layers, number of nodes in hidden layer; the 'k' parameter in k-means; etc.). Use Azure ML Studio only (either classic or new version). You may attach R or Python scripts.**

**Background of Problem**

A heart attack, as defined by the *National Institute on Aging*, occurs when the coronary arteries are not getting sufficient flow of oxygen-rich blood to the heart. If oxygen-rich blood is not restored to the heart quickly enough, either through medicine or a catheter, the heart muscles that are oxygen deprived will begin to die. Symptoms of a heart attack include chest pain or pressure and discomfort in the upper body, neck or arms, nausea, cold sweat, light-headedness or fainting, and shortness of breath. Statistics prove that each year, more than one million people in the United States have a heart attack and half of those individuals will die within an hour of experiencing symptoms before arriving at the hospital for treatment. Several factors can cause a heart attack. Such medical factors include high blood pressure, high blood cholesterol, smoking, diabetes, obesity, unhealthy diet, lethargy, and excessive alcoholic consumption. Victims of a heart attack can recover by undergoing what is called cardiac rehabilitation which include performing physical activity, being educated about healthy lifestyles, taking prescribed medication, refraining from activities that potentially caused the heart attack, and undergoing counselling in order to learn how to manage stress and to improve mental health.

**Motivation for Solving the Problem**

As mentioned previously, a heart attack could lead to death, so why should we wait until an individual experiences a heart attack to take action? There are measures to prevent a heart attack from happening such as living a healthy lifestyle, having a balanced work/life schedule to reduce stress, taking medication, however, even healthy individuals can experience a heart attack. Our goal is to determine who is more likely to experiencing a heart attack. We believe that being able to determine who is more prone to experiencing a heart attack, precautions could be taken to try and prevent more people from suffering from heart attacks. Such precautions include visiting a doctor to determine what measures to take depending on how severe his/her conditions are.

Do we know who is more prone to experiencing a heart attack? Hopefully, with the Heart Attack Analysis & Prediction Dataset from Kaggle, we will be able to determine which characteristics an individual has whether he/she is more prone to experiencing a heart attack. This dataset is beneficial because it is a relatively small dataset with fourteen columns. However, it only contains three hundred three rows which is not sufficient enough to be a true representative of the population. The columns include age, sex, chest pain (cp), resting blood pressure (trtbps), cholestoral (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalachh), exercise induced angina (exng), (oldpeak), (slp), (caa), (thall), and chance of experiencing a heart attack (output). Several of these measures will assist us in determining whether an individual is prone to experiencing a heart attack.

There are a variety of algorithms available, including Decision Trees, Naive Bayes, Logistic Regression, K-Nearest Neighbours, and Support Vector Machine, because the data being evaluated for model development. We are comparing the accuracies of decision tree and logistic regression when developing models.

Due to their simplicity and dependability, Two-Class Boosted Decision Trees and Two-Class Logistic Regression models were chosen as the two algorithms for the analysis. When properly set, both algorithms are capable of performing well on a range of machine learning tasks, predict an outcome variable in a supervised manner, and are optimized for binary variables like the chance a person will experience a heart attack (0 = less chance of getting a heart attack; 1 = more chance of getting a heart attack)

In Boosted Decision Tree, the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the second tree, and so forth, which is an ensemble learning technique. Decision trees have the disadvantage of being memory-intensive, making smaller datasets a better fit. The settings for a decision tree include a single parameter or range, the maximum number of leaves per tree (the value can improve precision but increases the risk of overfitting and lengthens the train time), the minimum number of samples per leaf node (with a default value of 1, a single case can create a rule), the learning rate (which defines the step size while learning and determines how quickly or slowly the learner progresses), and the minimum number of samples per leaf node. The number of trees constructed is the total number of trees to be created in the ensemble (more trees can mean better coverage but increased training time), the step size determines whether the learner will overshoot the solution (step size too large learner will do so, step size too small training takes longer to converge on the best solution), and the random number seed ensures reproducibility across runs with the same data and parameters.

It is common in classification use-cases to predict the outcome using a two-class logistic regression. The technique fits data to a logistic function to estimate the likelihood that an event will occur. This supervised learning technique trains the model using an outcome variable. The configuration for this algorithm includes a single parameter or a parameter range, optimization tolerance (which specifies a threshold value to use when optimizing), if the improvement between iterations falls below threshold the solution is reached and training stops, L1 and L2 regularization weight values are typically non-zero, and is a method to prevent overfitting by penalizing models with extreme coefficient values by adding the penalty to the error of the hypothesis.

-------

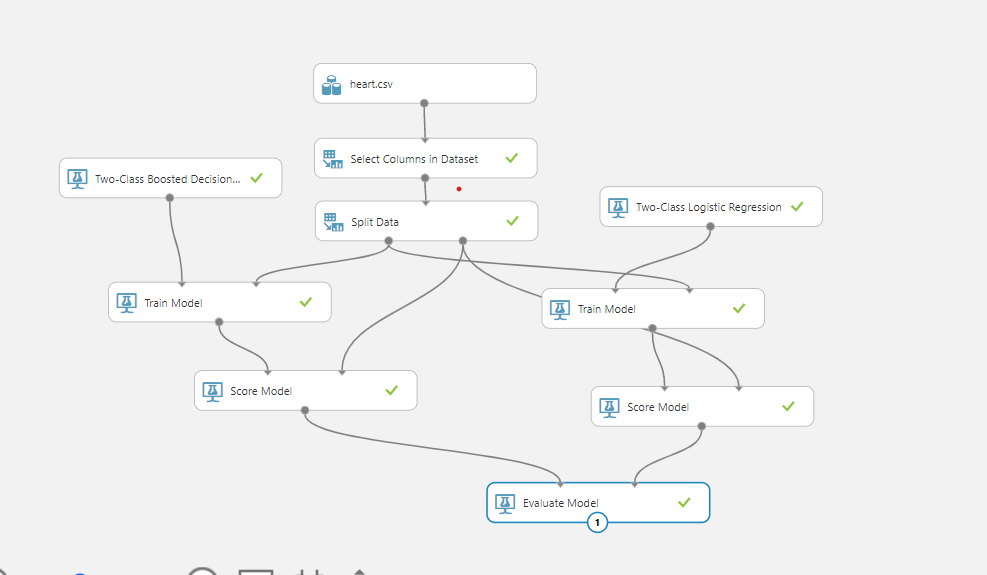
Decision trees are a decision-making tool that are formed in a manner similar to a flowchart. Another definition of a decision tree is a visual representation of a choice and every result that is feasible. Both continuous and categorical outputs can be produced using decision trees, where the criteria are represented by the decision nodes and the outcomes by the end nodes.

---

**Summary sheet showing the results of all experiments.**

By applying the first approach, the accuracy achieved by the Two-Class Boosted Decision Tree is 78.3%, Logistic Regression is 80.3. Logistic regression is having the highest accuracy here which is achieved by using the cross-validation and grid search for finding the best parameters or in other words doing the hyperparameter tuning. description abt Confusion matrix, the Logistic regression model is predicting True positives as 34 and True negatives as 15. Coming to Decision tree is predicting True positives as 33 and True negatives as 15.

* Attach the results here

****

**Chart

Description automatically generated**

**Graphical user interface

Description automatically generated with low confidence**

**Chart

Description automatically generated**

**A picture containing graphical user interface

Description automatically generated**

**Conclusions, including any recommendations for the company that is seeking to implement the solution.**

* Deep learning model can do better. We currently focusing on ML Algorithms