Experiment 1 - Identification of problem & creation of problem statement for selected application

**Aim:** To identify and search problem statement.

**Objective:**

The objective is to identifying and search the problem statement that can be solved by the machine learning algorithms.

**Theory:**

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you’re born with (congenital heart defects), among others.

The term “heart disease” is often used interchangeably with the term “cardiovascular disease”. Cardiovascular disease 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.

Heart disease is one of the biggest cause for morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subject in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decision and prediction.

This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Due to such constraints scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Here I will use ‘Heart Disease prediction dataset’ obtained from Kaggle/uci repository.

This Dataset has data of patients in Cleveland.

Database:

Data Source: Kaggle.com( <https://www.kaggle.com/ronitf/heart-disease-uci>)

Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1) from absence (value 0).

Attributes of Data:

1. **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.
2. **Sex** : Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman’s risk is similar to a man’s although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.
3. **Angina (Chest Pain)** : Angina is chest pain or discomfort caused when your heart muscle doesn’t get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.
4. **Resting Blood Pressure** : Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.
5. **Serum Cholestrol** : A high level of low-density lipoprotein (LDL) cholesterol (the “bad” cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the “good” cholesterol) lowers your risk of heart attack.
6. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body’s blood sugar levels to rise, increasing your risk of heart attack.
7. **Resting ECG** : For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.
8. **Max heart rate achieved**: The increase in the cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
9. **Exercise induced angina**: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the centre of your chest, but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinzmetal) Angina d. Microvascular Angina.
10. **Peak exercise ST segment** : A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an ‘equivocal’ test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischaemia); these patients are frequently referred urgently for coronary angiography.

Based on above features I will be performing data analysis, feature selection, and statistical modelling.

**Tools Proposed:**

* Numpy
* Pandas
* ScikitLearn
* Matplotlib
* Seaborn

**Conclusion:**

Thus we were able to source the data and get the correct definition of the data.

Experiment 2 - Requirement gathering & analysis

**Aim:**

To perform requirement gathering and analysis for the given dataset

**Objective:**

The prime objective of this experiment is to understand the data before initiating the EDA, statistical analysis and statistical modelling.

**Theory:**

Dataset Name: Cleveland.csv

Cleveland: No. of rows 303

Number of Attributes: 14 (including the predicted attribute)

Attribute Information:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type

* Value 1: typical angina
* Value 2: atypical angina
* Value 3: non-anginal pain
* Value 4: asymptomatic

1. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
2. chol: serum cholestoral in mg/dl
3. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
4. restecg: 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

1. thalach: maximum heart rate achieved
2. exang: exercise induced angina (1 = yes; 0 = no)
3. slope: the slope of the peak exercise ST segment

* Value 1: upsloping
* Value 2: flat
* Value 3: downsloping

1. Oldpeak: ST depression induced by exercise relative to rest
2. ca: number of major vessels (0-3) colored by flourosopy
3. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
4. target: diagnosis of heart disease (angiographic disease status)

* Value 0: < 50% diameter narrowing
* Value 1: > 50% diameter narrowing

**Goal:** Create a model that can predict patient has heart disease or not

**Conclusion:**

So in this experiment I was able to explore the data and data dictionary in domain perspective.

Experiment 3 – Designing and Implementation

**Aim:**

To perform EDA, Statistical Analysis and Statistical Modelling on the given dataset

**Objective:**

The objective of this experiment is to perform data analysis and modelling on Heart Dieses dataset to predict presence of heart disease in patient.

**Theory:**

In this experiment we will be performing the following statistical analysis:

1. Data Sanity Checks
2. Correlation analysis
3. Data Distribution Check
4. Outlier Detection and Removal
5. Separation of Features and Labels
6. Missing Value Analysis
7. Categorical Data Analysis and Dummy Encoding
8. Data Standardization

Following are the steps I will be performing for Data Modelling

1. Model Building using Logistic Regression
2. Model Building using K Nearest Neighbors Classification
3. Model Building using Support Vector Machine
4. Model Building using Decision Tree Classification
5. Model Building using Random Forest Classification

**Program:** Attached

**Conclusion:**

In this experiment, I was able to build Logistic Regression classifier with test accuracy of 90% for heart disease prediction.

Experiment 4 – Results and Performance

**Aim:**

To evaluate and get the best models based on the statistical and domain perspective.

**Objective:**

By the end of this experiment, we will be able to finalize the best model for performing prediction of presence of heart disease in patient based on cholesterol, ECG, etc features.

**Theory:**

A confusion matrix is a summary of prediction results on a classification problem.  
The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.  
The confusion matrix shows the ways in which your classification model is confused when it makes predictions.  
It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

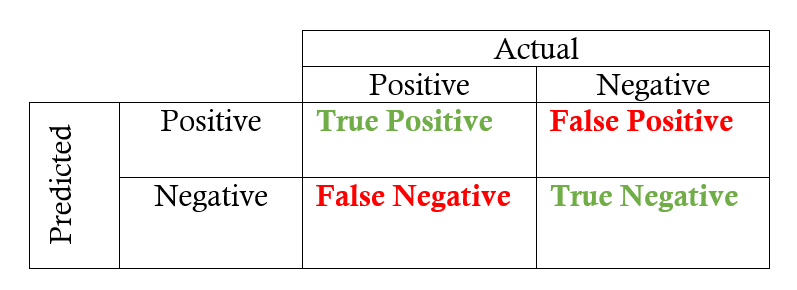


Fig. Confusion Matrix

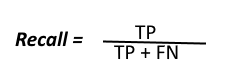
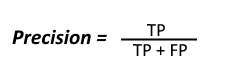
Here,  
• Class 1 : Positive  
• Class 2 : Negative

**Definition of the Terms:**  
• Positive (P) : Observation is positive (for example: is an apple).  
• Negative (N) : Observation is not positive (for example: is not an apple).  
• True Positive (TP) : Observation is positive, and is predicted to be positive.  
• False Negative (FN) : Observation is positive, but is predicted negative.  
• True Negative (TN) : Observation is negative, and is predicted to be negative.  
• False Positive (FP) : Observation is negative, but is predicted positive.

**Classification Rate/Accuracy:**  
Classification Rate or Accuracy is given by the relation:  

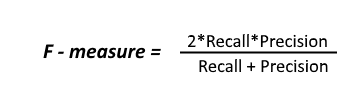

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

**Recall:**  
Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:  
  
**Precision:**  
To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).  
Precision is given by the relation:  


**High recall, low precision:**This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

**Low recall, high precision:**This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

**F-measure:**  
Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.  
The F-Measure will always be nearer to the smaller value of Precision or Recall.  


**Observations:**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr No. | Algorithm | Test Accuracy | Train Accuracy |
| 1 | Logistic Regression | 90.16% | 87.60% |
| 2. | KNN Classification | 88.52% | 87.60% |
| 3. | Support Vector Machine Classification | 86.88% | 87.19% |
| 4. | Decision Tree Classification | 81.96% | 100.0% |
| 5. | Random Forest Classification | 78.68% | 99.58% |

**Conclusion:**

Thus here we finalized the Logistic regression as our final deployable model for this experiment.

Experiment 5 – Limitations

**Aim:**

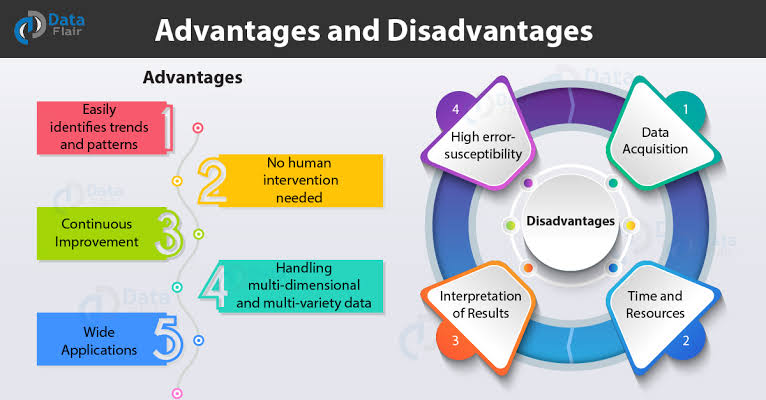
To understand and identify the current limitations of our application

**Objective:**

To explore the current limitation and propose a future solution to overcome the limitation

**Theory:**

Machine learning has revolutionized the world as we know it in the past decade. The information explosion has resulted in the collection of massive amounts of data. This amount of data, coupled with the rapid development of processor power and computer parallelization, has now made it possible to obtain and study huge amounts of data with relative ease.



# Limitation 1- Deterministic Problems:

# Machine learning is incredibly powerful for sensors and can be used to help calibrate and correct sensors when connected to other sensors measuring environmental variables such as temperature, pressure, and humidity. **Machine learning is stochastic, not deterministic.**

**Limitation 2- Data:**

This is the most obvious limitation. If you feed a model poorly, then it will only give you poor results. This can manifest itself in two ways: lack of data, and lack of good data.

**Lack of Data:**

Many machine learning algorithms require large amounts of data before they begin to give useful results.  Reusing data is a bad idea, and data augmentation is useful to some extent, but having more data is always the preferred solution. The larger the architecture, the more data is needed to produce viable results.

**Lack of Good data:**

Having a lack of good features can cause your algorithm to perform poorly; having a lack of good ground truth data can also limit the capabilities of your model.

# ****Limitation 3 - Interpretability:****

Interpretability is one of the primary problems with machine learning. An AI consultancy firm trying to pitch to a firm that only uses traditional statistical methods can be stopped dead if they do not see the model as interpretable.

These models as such can be rendered powerless unless they can be interpreted, and the process of human interpretation follows rules that go well beyond technical prowess. For this reason, interpretability is a paramount quality that machine learning methods should aim to achieve if they are to be applied in practice.

* No design guidelines. Accuracy depends on training and learning which is not always available.
* Hard to maintain degree of meaningfulness.
* Hard to combine cases together. Predictions are limited to the cases that have been observed.
* Required accurate details on many past projects.
* Have large data requirement to learn about various topics which may be time taking and cause various resources.

**Conclusion:**

In this experiment we were able to explore many limitations and solutions to overcome the limitation.