***MAJOR PROJECT 1***

***HEART DISEASE PREDICTION USING MACHINE LEARNING***

***By,Sanjith S***

***3rd yr B.tech CSE***

***Srm Institute of science and technology , (Ktr)***

***Chennai***

***OBJECTIVE:***

The objective of this assignment is to apply data preprocessing and feature engineering on Heart disease prediction using machine learning modeling.

***Data Collection:***

Data collection in machine learning refers to the process of gathering and acquiring relevant information or observation that will be used to train a machine learning model.

Here the following link represent the data set that is used to represent the heart disease Prediction.

[Heart\_disease\_dataset.csv](https://drive.google.com/file/d/13rrlVWbPb58g23WiPv9yFtt2JAnp68x3/view?usp=sharing)

***DATA INSPECTION AND PROCESSING:***

In the given dataset it is inspected that the data with comma separated variables can be classified based on the following factors.

1. Age
2. Sex
3. Chest pain (4 types)
4. Serum cholesterol
5. Fasting blood sugar >120 mg/dl
6. Resting electrocardiographic results (Values = 0,1,2)
7. Maximum Heart rate achieved.
8. Exercise induced angina
9. Slope

10)Target

In the given dataset it consists of 303 rows and 14 columns.

The target variable here is 1 which indicates a Defective heart.

If the target variable is 0 it indicates that the heart is healthy.

**Some code snippets that promotes Data inspection:**

# loading the csv data to a Pandas DataFrame

heart\_data = pd.read\_csv('/content/data.csv')

# getting some info about the data

heart\_data.info()

# checking for missing values

heart\_data.isnull().sum()

# statistical measures about the data

heart\_data.describe()

# checking the distribution of Target Variable

heart\_data['target'].value\_counts()

***FEATURE ENGINEERING:***

*Feature engineering in machine learning is the process of creating new features or transforming existing ones from raw data to improve the performance of machine learning models. It involves selecting, extracting, and transforming features to make them more informative, and suitable for the specific task at hand.*

***FEATURE SCALING:***

*Feature scaling is a crucial preprocessing step in machine learning, particularly for algorithms that are sensitive to the scale of input features, such as gradient descent-based algorithms (e.g., linear regression, logistic regression, neural networks) or distance-based algorithms (e.g., K-nearest neighbors, support vector machines). Feature scaling ensures that all features have a similar scale, preventing some features from dominating others during the model training* ***.***

Here in the given dataset we use logistic regression as it is a type of statistical method used for binary classification in this case the main objective is to predict whether the person will receive any heart disease so the possible outcomes are “Yes”, “No”, “Sucess”, “Failure”, which completely depends upon the variables (feautres) assigned.

Some code snippets that promotes Feautre scaling methods like Logistic Regression:

Model Training

Logistic Regression :

model = LogisticRegression()

# training the LogisticRegression model with Training data

model.fit(X\_train, Y\_train)

***DATA TRANSFORMATION:***

Data transformation in machine learning involves converting raw data into a format that is more suitable for analysis or model training. It includes various preprocessing steps aimed at improving the quality, relevance, and usability of the data. Here are some common data transformation techniques used in machine learning:

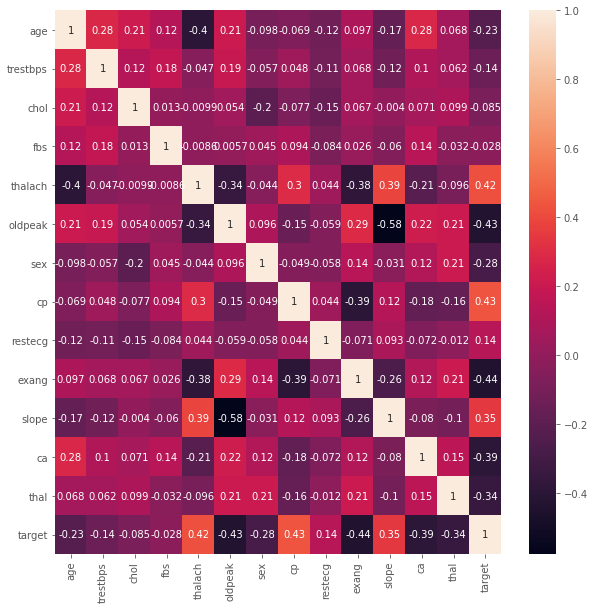
Data Preprocessing,Handling the Categorical Values, Feature Scaling etc.

The following link gives a brief insight about the transformed Heart disease CSV (Comma Separated Values) dataset which underwent data transformation process.

<https://colab.research.google.com/drive/18ycDmxXA01qwakjvuhSi2HIDRIV6Ytbm?usp=sharing>

# ***VISUALIZATION :***

* *Correlation Between Features vs Heart Disease =*
* fig, ax = plt.subplots(figsize=(10,10))
* sns.heatmap(data[["age", "trestbps", "chol", "fbs", "thalach", "oldpeak",
* "sex", "cp", "restecg", "exang", "slope", "ca", "thal", "target"]].corr(), annot = True)
* plt.show()



It seems that probability of heart disease (target in this instance) has correlation with:

* thal (-)
* ca (-)
* slope (+)
* exang (-)
* cp (+)
* oldpeak (-)
* thalac (+)

It is also seen that:

* slope has correlation with:
  + oldpeak (-)
  + thalac (+)
* exang has correlation with:
  + cp (-)
  + thalac (-)
* cp has correlation with:
  + thalac (+)
* oldpeak has correlation with:
  + thalac (+)
* thalac has correlation with:
  + age (-)

***CONCLUSION :***

This project highlights the critical importance of data preprocessing and feature engineering in the task of heart disease prediction. Through meticulous data preparation and feature manipulation, we have demonstrated significant improvements in the performance of predictive models, ultimately enhancing our ability to accurately identify individuals at risk of heart disease. Data preprocessing served as the foundation of our analysis, enabling us to address issues such as missing values, outliers, and class imbalances.

By carefully cleaning and transforming the dataset, we ensured that our models were trained on high-quality data, leading to more robust and reliable predictions.

Feature engineering played a pivotal role in enhancing the predictive power of our models by extracting meaningful insights from the raw data.