**Predicting Heart Disease Using Machine learning**

**PROBLEM:** Given clinical parameters about a patient, can we predict whether or not they have heart disease.

**DATASET:**

* **heart.csv**

**DATA ANALYSIS:**

**1. age - age in years**

**2. sex - (1 = male; 0 = female)**

**3. cp - chest pain type**

**\* 0: Typical angina: chest pain related decrease blood supply to the heart**

**\* 1: Atypical angina: chest pain not related to heart**

**\* 2: Non-anginal pain: typically esophageal spasms (non heart related)**

**\* 3: Asymptomatic: chest pain not showing signs of disease**

**4. trestbps - resting blood pressure anything above 130-140 is typically cause for concern**

**5. chol - serum cholestoral**

**\* serum = LDL + HDL + .2 \* triglycerides**

**[LDL : Low density lipoproteins, HDL: High density lipoproteins]**

**\* above 200 is cause for concern**

**6. fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)**

**\* “>126” mg/dL signals diabetes**

**7. restecg - resting electrocardiographic results**

**\* 0: Nothing to note**

**\* 1: ST-T Wave abnormality can range from mild symptoms to severe problems**

**\* signals non-normal heart beat**

**2: Possible or definite left ventricular hypertrophy**

**Enlarged heart's main pumping chamber**

**8. thalach - maximum heart rate achieved (220-Age)**

**9. exang- exercise induced angina(type of chest pain)**

**(1 = yes; 0 = no) [Breathing Problem]**

**10. oldpeak- ST depression induced by exercise relative to rest looks at stress of heart during excerise unhealthy heart will stress more**

**11. slope - the slope of the peak exercise ST segment**

**\* 0: Upsloping: better heart rate with excercise (uncommon)**

**\* 1: Flatsloping: minimal change (typical healthy heart)**

**\* 2: Downslopins: signs of unhealthy heart**

**12. ca - number of major vessels (0-3) colored by flourosopy**

**\* colored vessel means the doctor can see the blood passing through**

**\* the more blood movement the better (no clots)**

**13. thal - thalium stress result**

**\* 1,3: normal**

**\* 6: fixed defect: used to be defect but ok now**

**\* 7: reversable defect: no proper blood movement when excercising**

**14. target - have disease or not (1=yes, 0=no)**

# Exploratory Data Analysis:

# Step1: Import modules

# Step2: Import data

# Step3: Perform basic operations on data by using Pandas

# Step-4: After performing operations by using Data Visualization:

* Now I want to know how many people have heart disease and how many people don't by using hvplot

Input: Target column in dataset

Output: We have 165 person with heart disease and 138 person without heart disease, so our problem is balanced

* Find missing values in dataset .Then I find there is no missing values.
* Now I divide my dataset into two parts as categorical values and continuous values then I create subplots for these values in the form of histograms.
* Now I find correlation between target column with other columns then I realized that

fbs and col  are the lowest correlated with the target variable.

All other variables have a significant correlation with the target variable.

* After exploring the dataset, I observed that I need to convert some categorical variables into dummy variables and scale all the values before training the Machine Learning models. First, I'll use the get\_dummmies method to create dummy columns for categorical variables.

MODEL BUILDING:

* Now I spilt data into X\_Train, X\_test ,Y\_Train , Y\_Test.
* Now I want to check quality of a prediction.
* Now I want to do the visualizes and summarizing the performance of classification problem
* Now I divide data into two parts such as training and testing data.
* We use some algorithms for get good effiencent performance.
* We're going to try 4 different machine learning models
* Logistic Regression
* K-Nearest Neighbours Classifier
* Support Vector machine
* Decision Tree Classifier
* Random forest
* XG Boost Classfier
* Gradient boost regress
* Naïve Bayes

1. LOGISTIC REGRESSION:

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of x.

A classification report shows the precision, recall, f1-score,support of classification model.

F1 score is to 1.0 the better the expected performance of the model. But I get 0.85 for training ,0.87 for testing ,this is not perform well .That’s why we move onto the K-Nearest Neighbor Algorithm.

\*\*\* For training data accuracy rate: 86.79% \*\*\*

\*\*\* For testing data accuracy rate: 86.81% \*\*\*

2) K-NEAREST NEIGHBOUR:

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

A classification report shows the precision, recall, f1-score, and support of classification model.

F1 score is to 1.0 the better the expected performance of the model. But I get 0.87 for training, 0.87 for testing this is not perform well .That’s why we move onto the Support Vector Machine Algorithm.

\*\*\* For training data accuracy rate: 86.79% \*\*\*

\*\*\* For testing data accuracy rate: 86.81% \*\*\*

1. SUPPORT VECTOR MACHINE:

A support vector machine (SVM) is a type of [deep learning](https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network) [algorithm](https://www.techtarget.com/whatis/definition/algorithm) that performs supervised learning for classification or regression of data groups.

In [AI](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence) and [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML), [supervised learning](https://www.techtarget.com/searchenterpriseai/definition/supervised-learning) systems provide both input and desired output data, which are labeled for classification. The classification provides a learning basis for future data processing. Support vector machines are used to sort two data groups by like classification. The algorithms draw lines (hyperplanes) to separate the groups according to patterns.

A classification report shows the precision, recall, f1-score, support of classification model.

F1 score is to 1.0 the better the expected performance of the model. But I get 0.93 for training, 0.88 for testing this is not that much of perform .That’s why we move onto the Decision Tree Algorithm.

\*\*\* For training data accuracy rate: 93.40%\*\*\*

\*\*\* For testing data accuracy rate: 87.91%\*\*\*

4) DECISION TREE ALGORITHM:

A decision tree is a graph that uses a branching method to illustrate every possible output for a specific input

A classification report shows the precision, recall, f1-score, Support of classification model.

F1 score is to 1.0 the better the expected performance of the model. I am also get 1.0 for training and 0.78% this is performing very well.

\*\*\* For training data accuracy rate: 100.00%\*\*\*

\*\*\* For testing data accuracy rate: 78.93%\*\*\*

5)RANDOM FOREST:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

6) XG BOOST CLASSFIER:

**XGBoost**, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.

7)GRADIENT BOOSTING REGRESSO:

**Gradient boosting Regression** calculates the difference between the current prediction and the known correct target value. This difference is called residual.

8) NAÏVE BAYES:

 It is **a classification technique based on Bayes' Theorem with an assumption of independence among predictors**. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

HYPERMETER TUNING:

Hyparmeter tuning is crucial as they control the overall behavior of the machine learning model.

1. Logistic Regressions:

* Here we importing module called as GridSearchCV for tuning.
* The learning rate for training a neural networks.

1. K-Nearest Neighbor:

* The KNN algorithm can complete with the most accurate models because it makes highly accurate predictions.

Therefore, you can create the KNN algorithm for applications that require high accuracy but that do not require a human-readable model .The quality of the predictions depends on the distance measure.

1. Support Vector Machine:

* SVM is a widely used supervised learning algorithm.It is used in classfications tasks.
* Tuning the parameters values for machine learning algorithms efficiently improves model performance.
* Two important parameters of support vector machines:
* C-Parameter
* GAMMA-Parameter

1. Decision Tree :

* Hyperparameters are arguments accepted by a model-making function and can be modified to reduce overfitting,leading to a better generalizations of the model.

NOTE:

* Before tuning we get 100% training accuracy in decision tree algorithm , because we are using a part of training data for testing.After the time of training,decision tree gained the knowlodge about the data and now if you give some data to predict tree producing correct results every time.
* After tuning there is a reduced overfitting accuracy.

-MODEL DEVELOPED BY:

D .Sai Sri Divya, S180559

Y. Chanti, S180277

# 

# 

# 