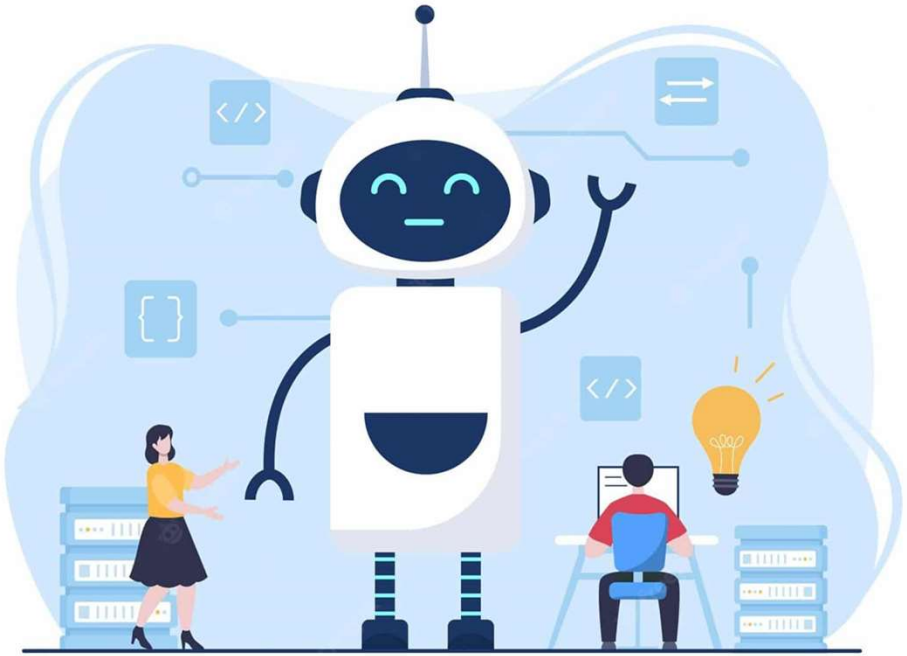
Modification of Naïve Bayes Classifier

(Research and Development in Machine Learning)

INTRODUCTION



My research project focuses on **enhancing Naïve Bayes Classifier** through innovative modification techniques. Through a blend of rigorous **research** and hands-on **development**, uncovering new possibilities in the realm of machine learning.

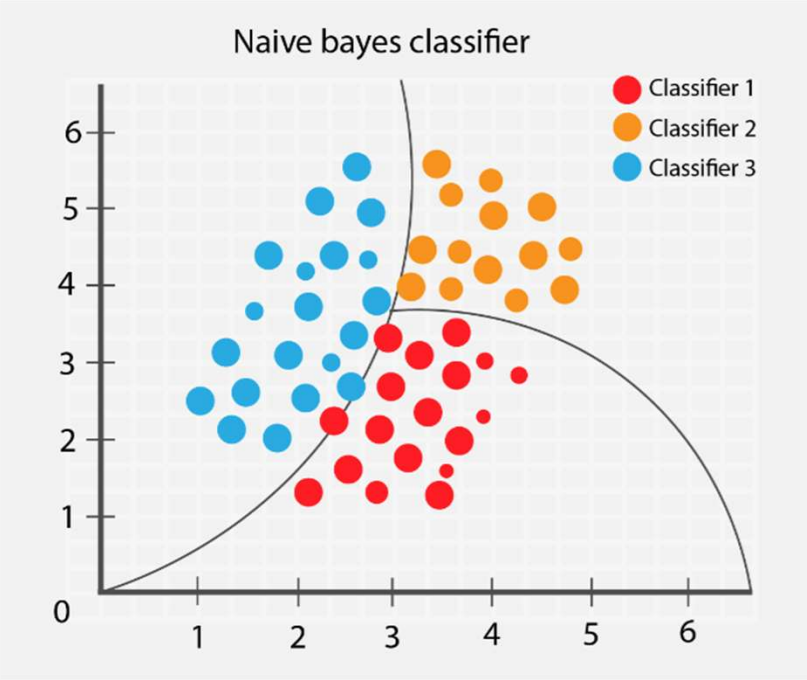
Naive Bayes classifier

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



Gaussian Naive Bayes

Gaussian Naive Bayes is a machine learning classification technique based on a probablistic approach that assumes each class follows a normal distribution. It assumes each parameter has an independent capacity of predicting the output variable. It is able to predict the probability of a dependent variable to be classified in each group.

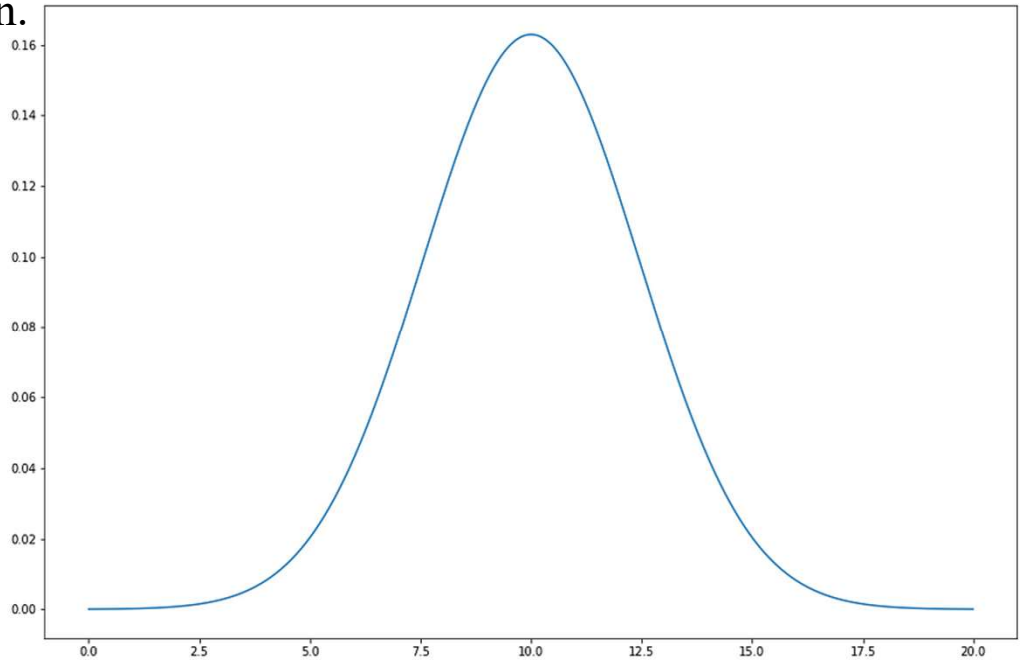
The combination of the prediction for all parameters is the final prediction that returns a probability of the dependent variable to be classified in each group. The final classification is assigned to the group with the higher probability.

Gaussian Distribution

Gaussian distribution is also called normal distribution. Normal distribution is a statistical model that describes the distributions of continuous random variables in nature and is defined by its bell-shaped curve. The two most important features of the normal distribution are the mean (μ) and standard deviation (σ). The mean is the average value of a distribution, and the standard deviation is the "width" of the distribution around the mean.

A variable (X) that is normally distributed is distributed continuously (continuous variable) from $-\infty < X < +\infty$, and the total area under the model curve is 1.

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



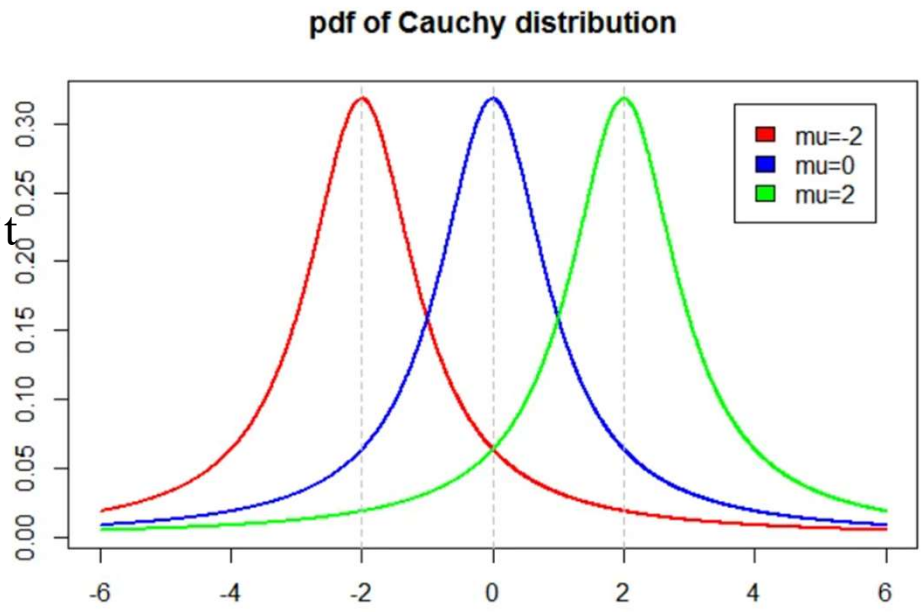
Cauchy Distribution

The Cauchy distribution, sometimes called the Lorentz distribution, is a family of continuous probably distributions which resemble the normal distribution family of curves. While the resemblance is there, it has a taller peak than a normal. And unlike the normal distribution, it's fat tails decay much more slowly.

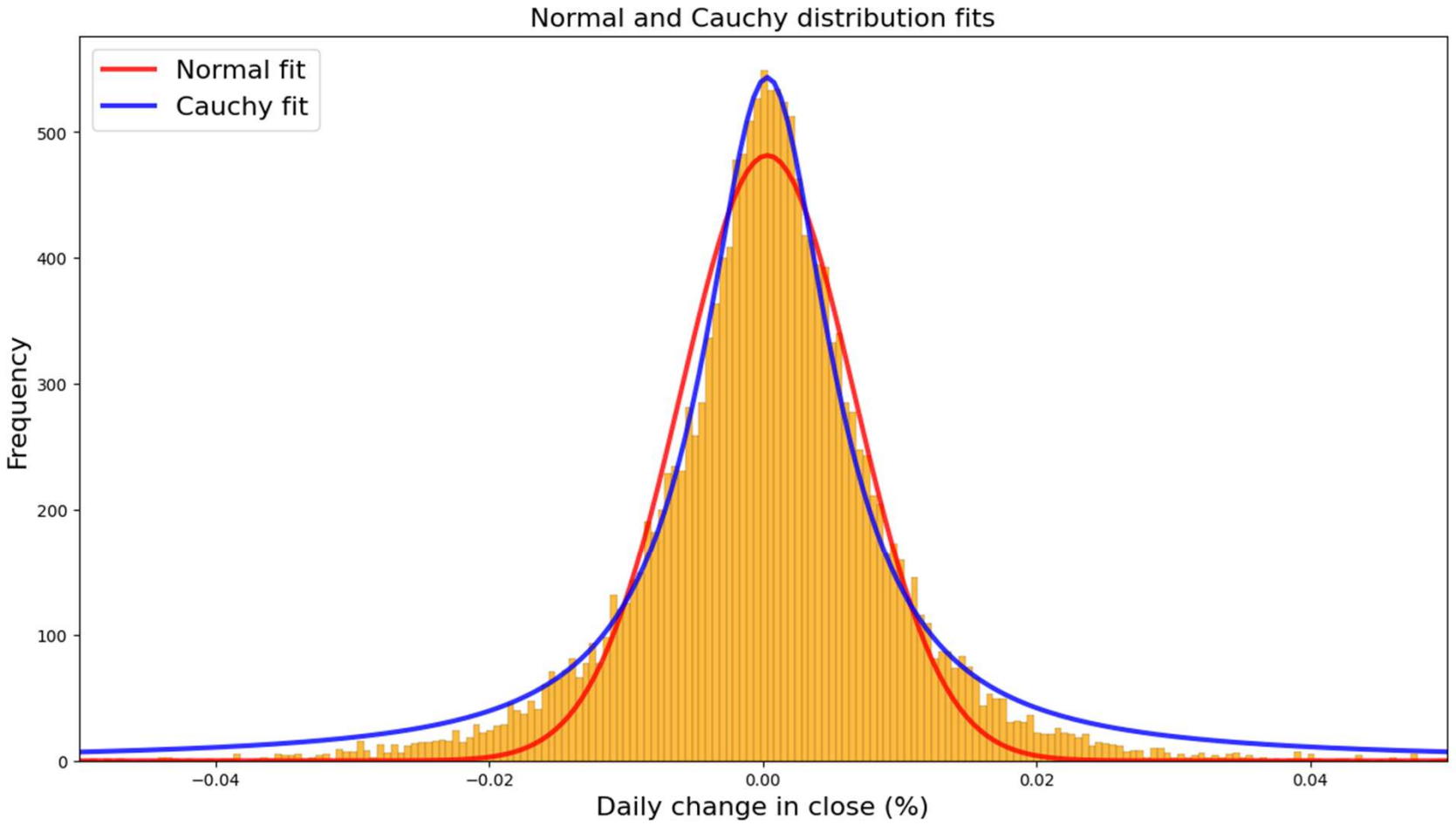
The Cauchy distribution is well known for the fact that it's expected value and other moments do not exist. The median and mode do exist. And for the Cauchy, they are equal. Together, they tell you where the line of symmetry is. However, the Central Limit theorem doesn't work for the limiting distribution of the mean. In sum, this distribution behaves so abnormally it's sometimes considered the Hannibal Lecter of distributions.

$$f(x) = \frac{1}{s\pi(1+((x-t)/s)^2)}$$

$$f(x) = \frac{1}{\pi(1+x^2)}$$



Gaussian Vs Cauchy

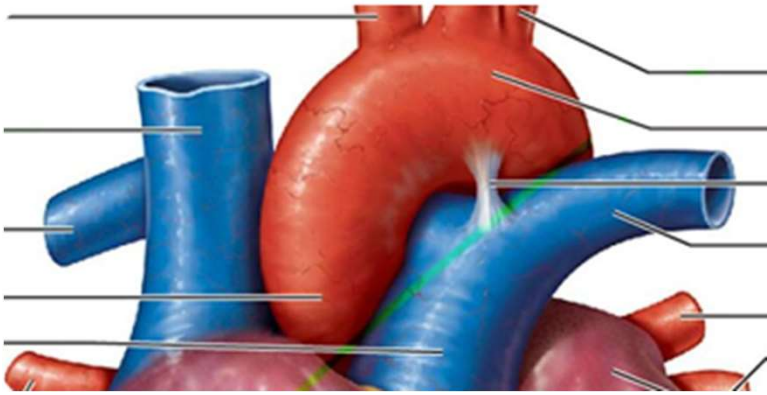


LITERATURE REVIEW

Sr No.	Paper Title	Contribution
1.	**Emotion Recognition Using a Cauchy Naive Bayes Classifier. (IEEE)**	In this paper a method proposed for recognizing emotions through facial expressions displayed in video sequences. We introduce the Cauchy Naive Bayes classifier which uses the Cauchy distribution as the model distribution, and provide a framework for choosing the best model distribution assumption. Our person-dependent and person-independent experiments show that the Cauchy distribution assumption typically provides better results than the Gaussian distribution assumption.
2.	**Circular Bayesian classifiers using wrapped Cauchy distributions. (ScienceDirect)**	The paper introduces four supervised Bayesian classification algorithms for circular variables, using wrapped Cauchy distributions. These classifiers leverage the unique properties of the bivariate wrapped Cauchy distribution. Synthetic data and a real neuromorphological dataset are employed to demonstrate and evaluate the proposed algorithms, which show promising predictive performance compared to several other classification methods including Gaussian TAN classifier, decision tree, random forest, multinomial logistic regression, support vector machine, and simple neural network.

DATASET

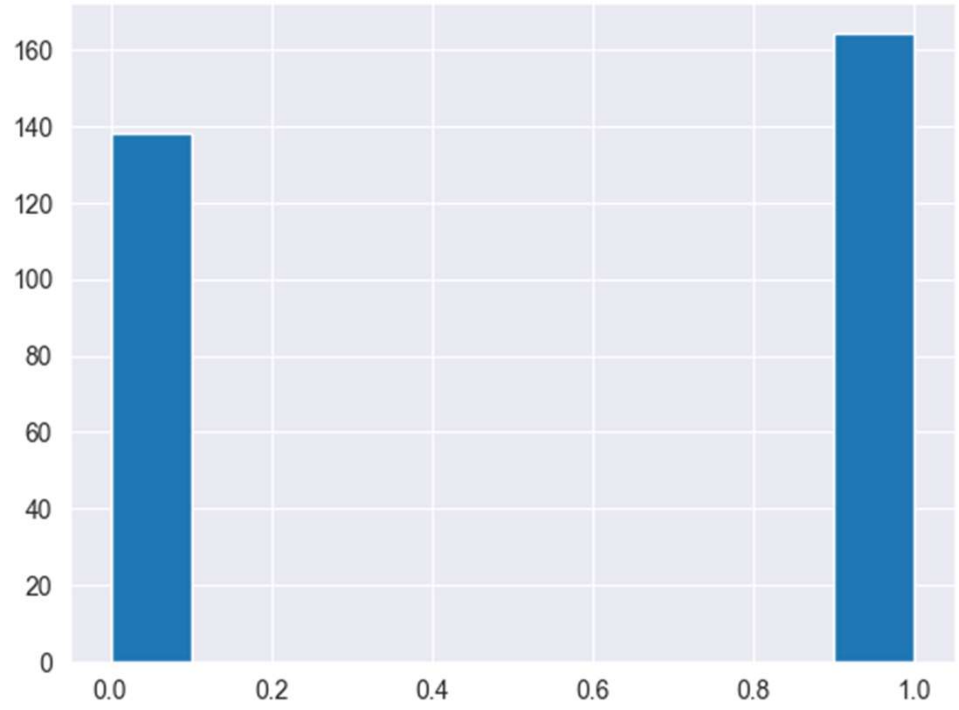
Heart Attack Analysis & Prediction Dataset

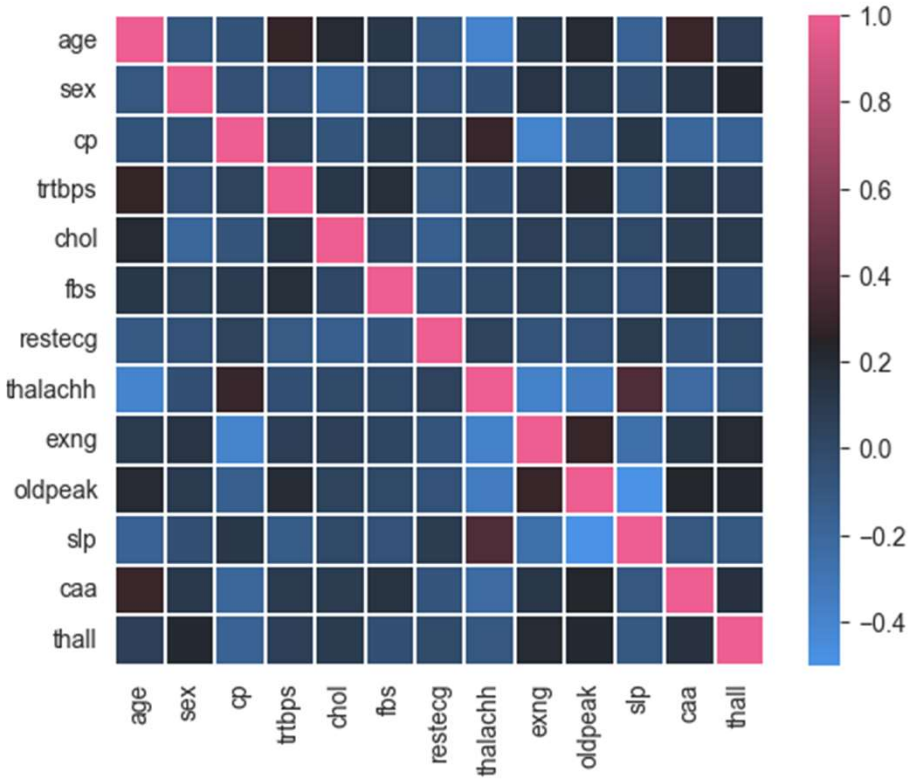


About this dataset

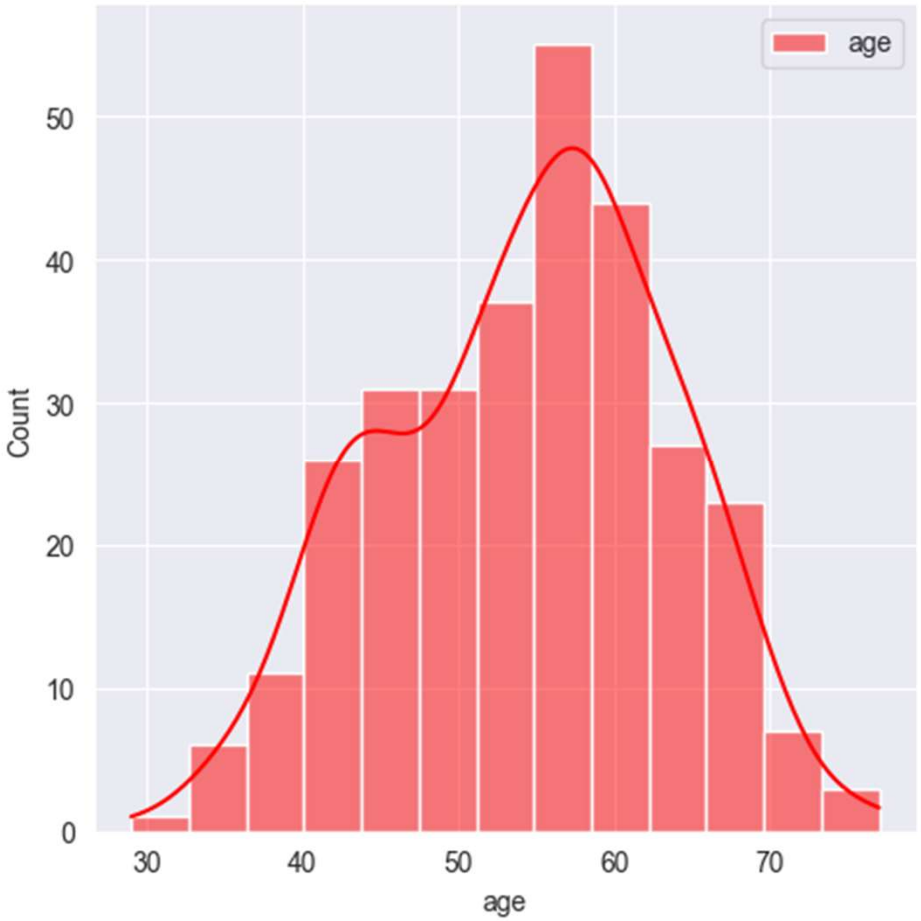
- Age : Age of the patient
- Sex : Sex of the patient
- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
- cp : Chest Pain type chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- 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

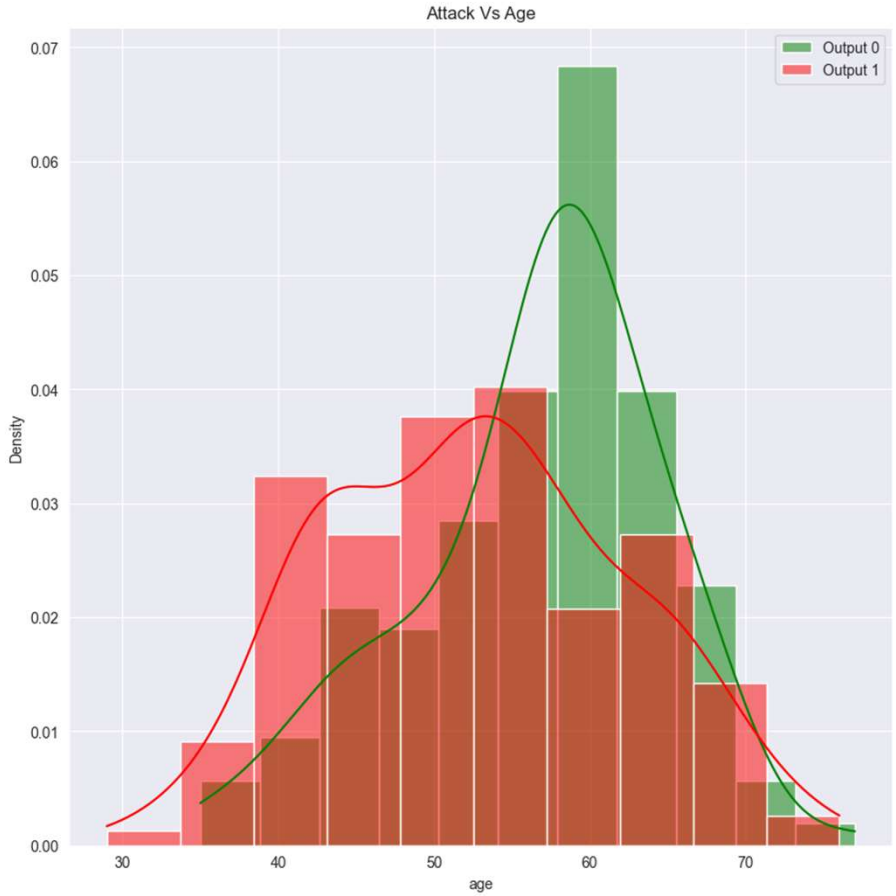
Visualization and Analysis



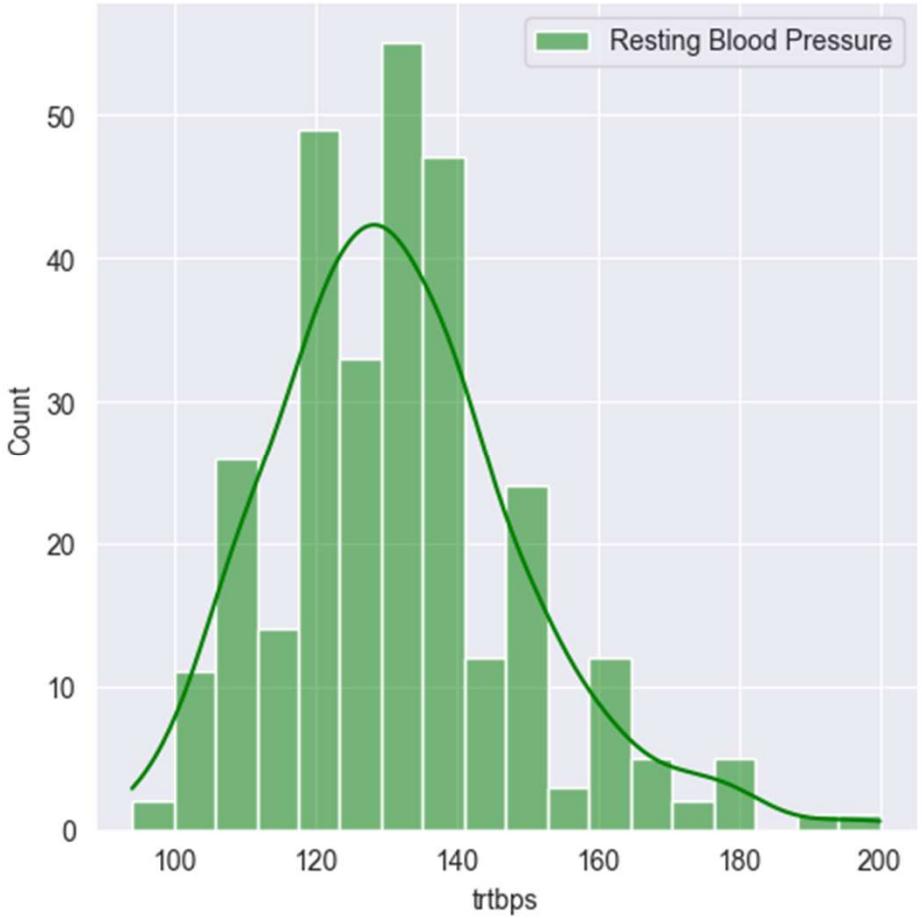


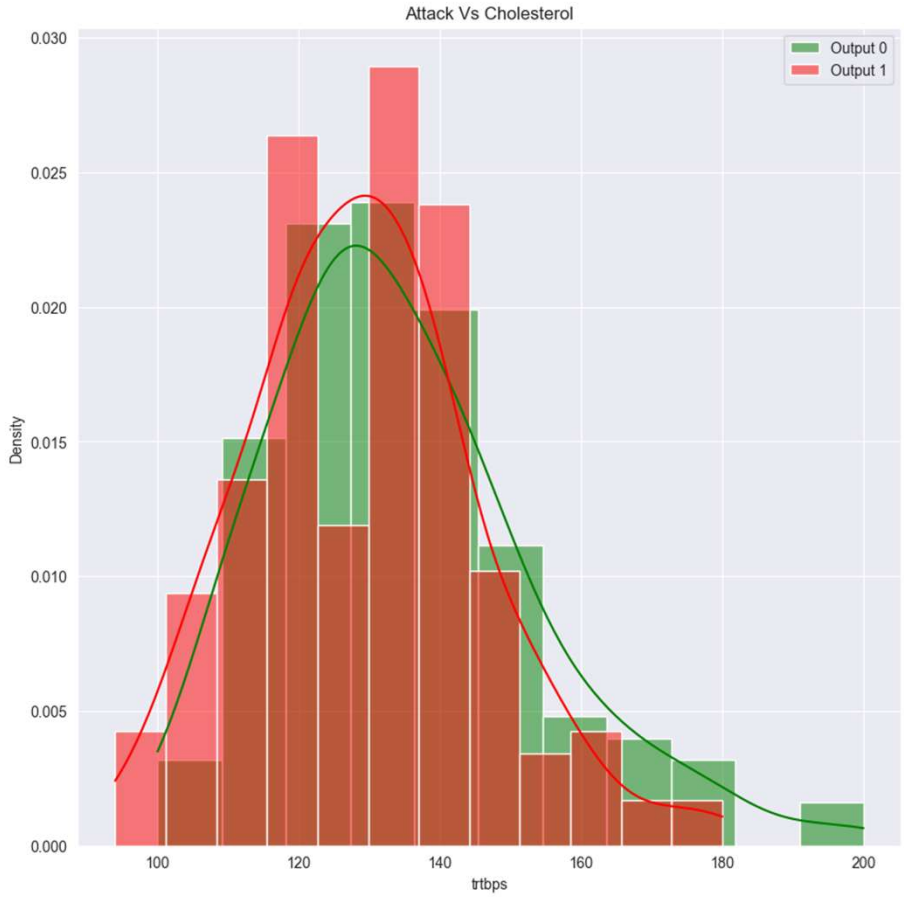
Visualization and Analysis



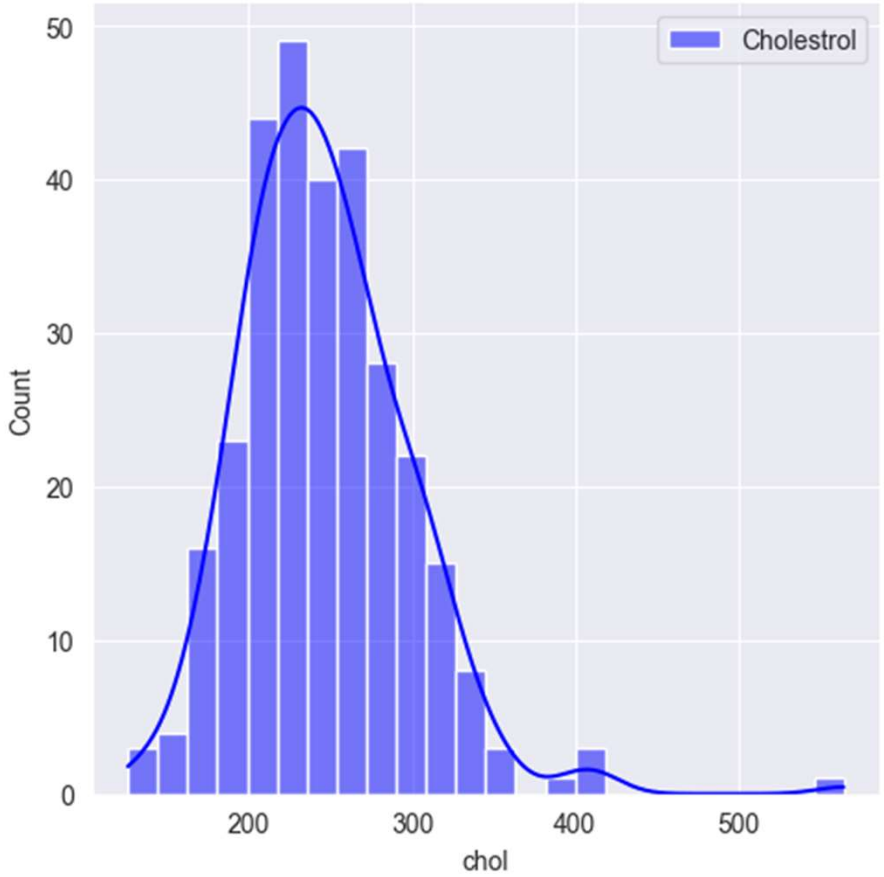


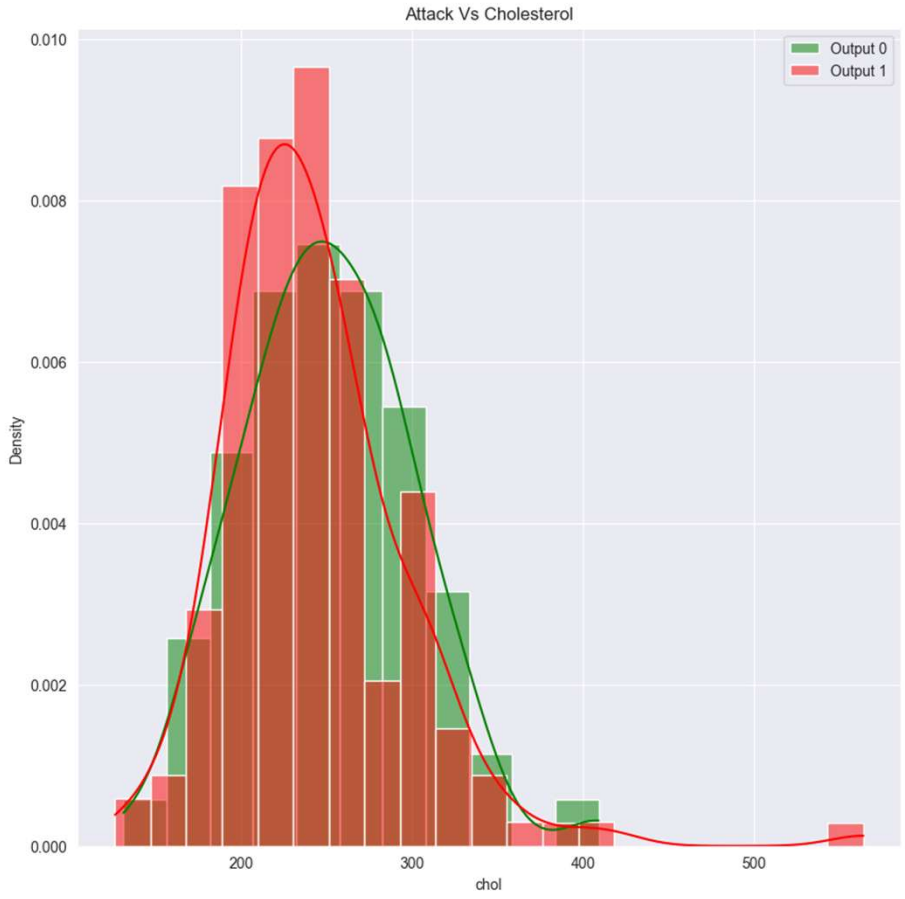
Visualization and Analysis





Visualization and Analysis





PROBLEM STATEMENT

STATEMENT: The Naive Bayes classifier used Gaussian Distribution for the classification. Modifying the distribution(use Cauchy distribution instead of Gaussian distribution) used in Naive Bayes and comparing its results with earlier used distribution technique.

Prediction

Gaussian Naive Bayes

	Test size=0.1	Test size=0.2	Test size=0.25	Test size=0.3
Precision	0.9333333333333333	0.896551724137931	0.8974358974358975	0.8888888888888888
Recall	0.8235294117647058	0.8125	0.8536585365853658	0.8163265306122449
Accuracy	0.8709677419354839	0.8524590163934426	0.868421052631579	0.8461538461538461
F1 Score	0.875	0.8524590163934426	0.875	0.851063829787234

Prediction

Cauchy Naive Bayes

	Test size=0.1	Test size=0.2	Test size=0.25	Test size=0.3
Precision	0.8823529411764706	0.8529411764705882	0.8636363636363636	0.8461538461538461
Recall	0.8823529411764706	0.90625	0.926829268292683	0.8979591836734694
Accuracy	0.8709677419354839	0.8688524590163934	0.881578947368421	0.8571428571428571
F1 Score	0.8823529411764706	0.8787878787878788	0.8941176470588236	0.8712871287128713

Comparison

	Gaussian Naive Bayes	Cauchy Naive Bayes
Precision	0.90405246094901	0.86127108185932
Recall	0.82650361974058	0.90334784828566
Accuracy	0.85950041427859	0.86963550136579
F1 Score	0.86338071154517	0.88163639893401

RESULT

Our experiment shows that the Cauchy distribution assumption typically provides better results than the Gaussian distribution assumption.