

Heart Disease Risk Prediction

Machine learning project

Mohamed Yunes (12331354) & Nour Saneh (12210255)

# Introduction

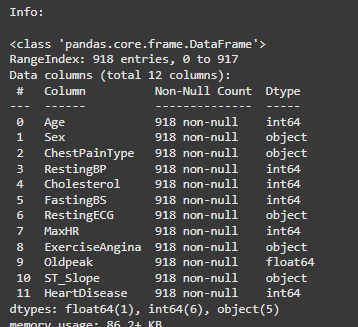
This report’s objective is to determine the best machine learning model for predicting heart disease in a patient. The report compares 4 different classification models trained on the same dataset using several metrics to choose the most reliable model.

Machine learning can analyze vast amounts of patient data to uncover patterns and risk factors that might not be apparent through traditional methods. In theory, this can aid healthcare providers in identifying high-risk individuals who might benefit from more intensive monitoring or lifestyle changes, and limit rates of preventable heart disease death. This can also help us understand and study the underlying reasons for heart disease.

The process for machine learning first starts with analyzing and cleaning the data, then training the models and evaluating their performance by comparing key metrics.

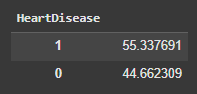
# Data Preprocessing

Before training the models, the data must first be prepared. We chose a dataset from Kaggle with 918 samples and 12 features related to heart health.



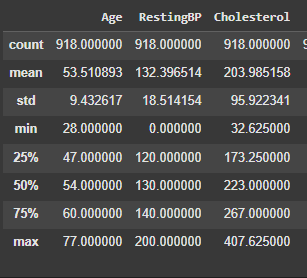
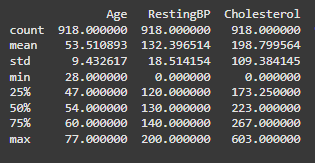
After loading the dataset in Python we made sure the data is clean and does not have any null or duplicated values. We also made sure no column has 0 standard deviation as it would be redundant in training.

While looking at the data we noticed that the target class division is unbalanced (55% positive to 45% negative), which will need to be addressed in evaluation.



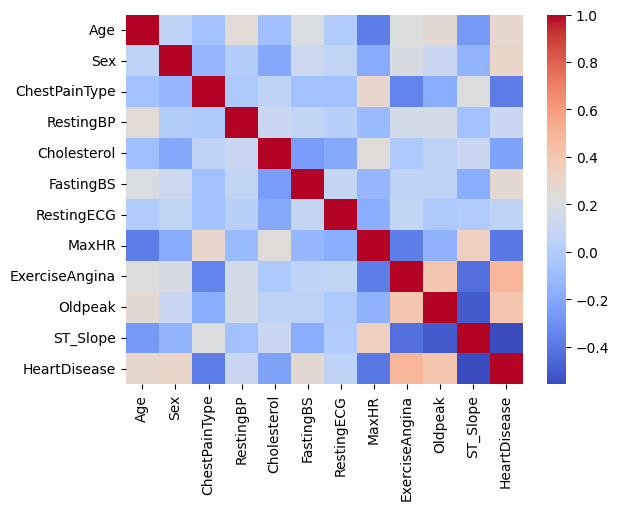
We also noticed that the gender was also unbalanced (78% male to 22% female) which will be permitted for this report but would be problematic in a real world application.

After looking at the standard deviations we also noticed that cholesterol had outliers so we fixed that using IQR:



Then we encoded the categorical features (sex, chest pain type, resting ecg, etc.) so the models could quantify them for training and testing.

Additionally, we made sure no columns highly correlate (using a correlation matrix) as they would also be redundant.



We isolated the Heart Disease column as a target and set the rest of the attributes as features.

As a final step before training the models, we split the data into a training set (20%) and testing set (80%) to be able to evaluate their performance. The data is then normalized so it can be used in training.

# Model Training & Evaluation

Our target is a binary so this is a classification problem. Then we have 4 potential models to choose from: Naïve Bayes, Logistic Regression, K-Nearest Neighbor and Random Forest. In order to decide which model is best suited we trained them on the same dataset and compared their metrics.

## Gaussian NB:

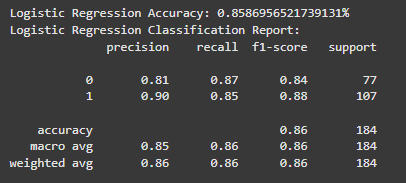
This model uses Bayes' theorem to make predictions, which is particularly effective for small datasets with normally distributed features. Our dataset matches these criteria so in theory NB should be an effective model.

After training the model these are the metrics recorded:

## Logistic Regression:

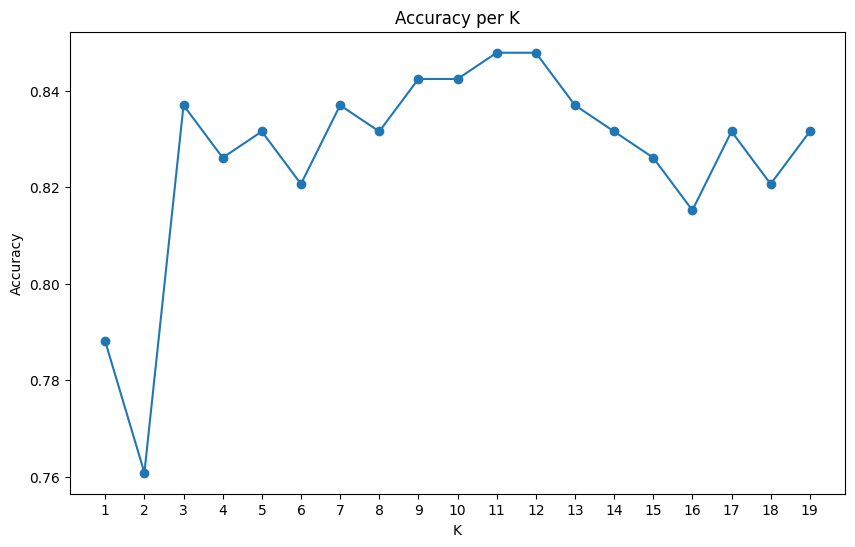
This is a linear model for binary classification that estimates the probability of a class label using a logistic function.

After training the model these are the metrics recorded:

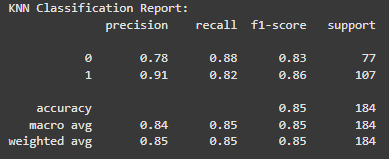


## KNN:

This algorithm classifies a sample based on the majority class of its K-nearest neighbors in the feature space. To maximize efficiency of the model we tested several values of K (1 to 20) to determine the value with the best accuracy result:

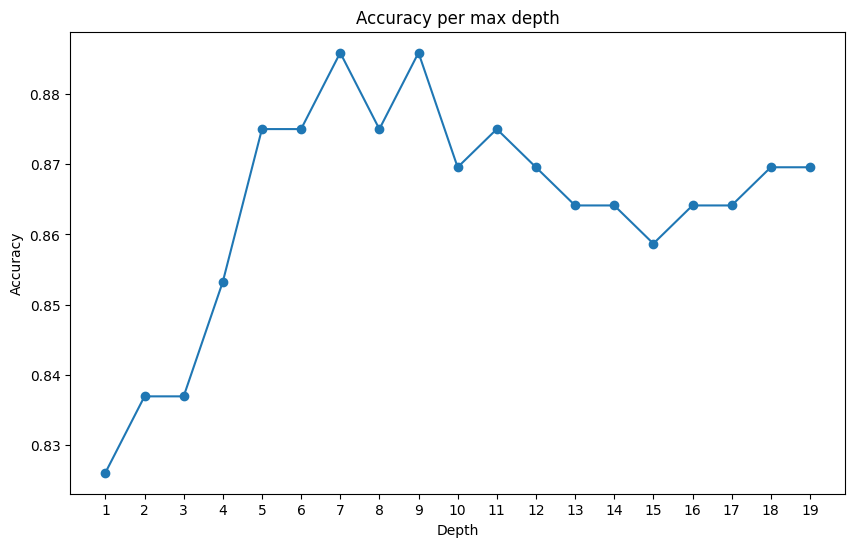


As per the graph we determined that the suitable K for our training will be K=11 resulting in 84.7% accuracy.

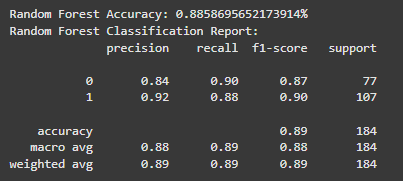


## Random Forest:

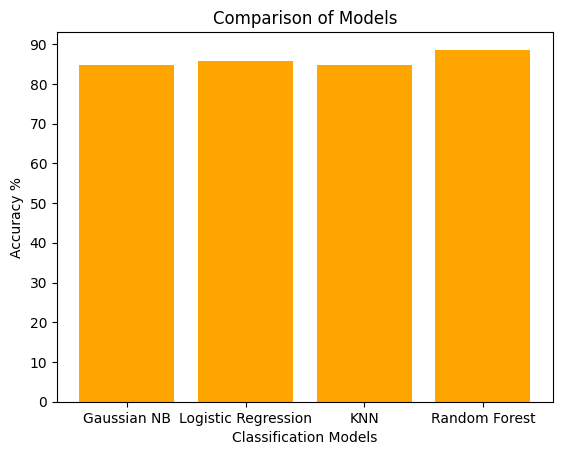
This model constructs multiple decision trees during training and outputs the mode of their predictions for classification. The advantage of this model is its resistance to overfitting even with the default depth. However, we still measured several depth values (1 to 20) to ensure the most accurate model:

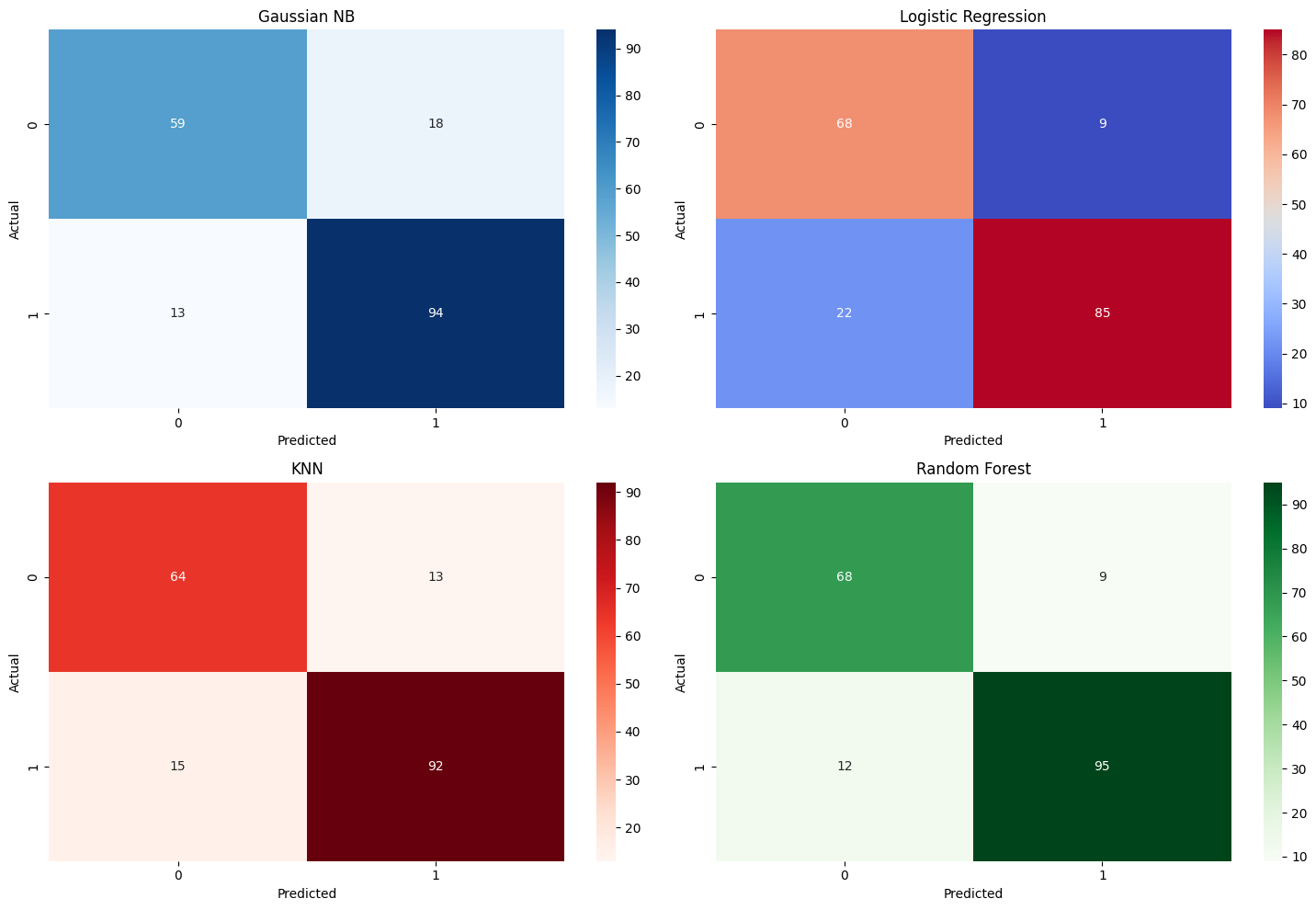


We determined that the most efficient depth value is 7, resulting in 88.5% accuracy.



Upon training and evaluating all of the models we compiled these graphs for the accuracies and confusion matrices:





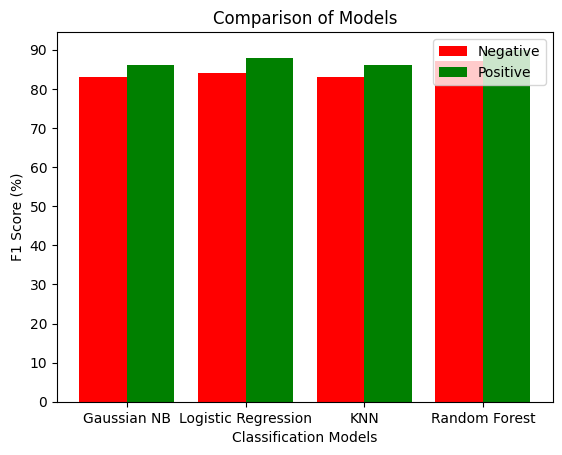
These results show that while all 4 models performed decently, random forest scored the best accuracy and most efficient confusion matrix, recording the highest number of true positives and true negatives, as well as lowest number of false negatives. The importance of this will be explained in the results section.

# Results

In order to decide which model was the most suitable we must compare the metrics in the evaluations. Previously we compared the accuracies which is a baseline comparison for all classification models. However there are other metrics to draw from, most importantly in the case of heart disease prediction the comparison of false negatives.

The presence of false negatives can prove to be very dangerous since it could allow patients to be ignored when they should be given attention. Thus the model chosen should have the lowest rate of false negatives. Consulting the confusion matrix, we concluded that random forest fits this criterion.

Another metric to consider is F1 score, which in this case is more important than accuracy as the dataset is not completely balanced as well as the aforementioned danger of false negatives.



Comparing the F1 scores of the 4 models, random forest again comes up as the winner.

All in all, the random forest performed the best of all the models, which makes sense as it is the least susceptible to both overfitting and outliers.

# Conclusion & Recommendation

After preparing the data and training the 4 classification models to predict heart disease affliction, we found that the random forest model had the most favorable metrics in the end. Machine learning can improve survival rates of patients and has the potential to prevent many deaths in healthcare.

However, the potential problems of using machine learning to predict heart disease risk include the disparity of available data between certain features like sex and also the still-present (although limited) risk of false negatives. This means machine learning might not be the perfect approach to an objective like this one.

With larger datasets it might be more suitable to alternatively use deep learning instead of machine learning as that would be more cost effective and more accurate.