**PGP in Data Science**



**HEART DISEASE PREDICTION**

**TECHNICAL ANALYSIS REPORT**

By:

Manikandan Raja (D22029)

# **Table of Contents**

[**Introduction** 5](#_Toc103121215)

[**Problem Statement** 6](#_Toc103121216)

[**Objective** 7](#_Toc103121217)

[**Methodology** 9](#_Toc103121217)

[**Exploratory Data Analysis** 17](#_Toc103121224)

[ **Checking for null values** 17](#_Toc103121224)

[ **Shapiro-Wilks Test** 17](#_Toc103121224)

[ **Checking for Duplication** 17](#_Toc103121224)

[ **Checking for Outliers** 17](#_Toc103121224)

[ **Correlation matrix** 17](#_Toc103121224)

[ **Checking Distribution using Histogram** 17](#_Toc103121224)

[**Fitting Baseline**  17](#_Toc103121224)

[ **Train-Test-Split** 17](#_Toc103121224)

[ **K-Fold CV………………………………………………………………………………** 17](#_Toc103121224)

[**Feature Selection Technique** 17](#_Toc103121224)

[ **ANOVA** 17](#_Toc103121224)

[ **Cramer -V test** 17](#_Toc103121224)

[ **Multicollinearity Check using VIF** 17](#_Toc103121224)

**[Models Used](#_Toc103121233)** [19](#_Toc103121233)

[ **Logistic Regression** 19](#_Toc103121234)

[ **DecisionTreeClassifier** 19](#_Toc103121235)

[ **RandomForestClassifier** 19](#_Toc103121235)

[ **AdaBoostClassifier** 19](#_Toc103121235)

[ **BaggingClassifier** 19](#_Toc103121235)

[ **GradientBoostingClassifier** 19](#_Toc103121235)

[ **StackingClassifier** 19](#_Toc103121235)

[**Models Performance after Hyperparametric Tuning** 20](#_Toc103121237)

[**Results** 20](#_Toc103121242)

[**Interpretation** 21](#_Toc103121243)

[**Conclusion** 30](#_Toc103121263)

**Link for Code and Datasets -** [**Google Drive Link**](https://drive.google.com/drive/folders/1XvpNsLbP_Eei9VDI0E5OEcuSpSJqR6ON?usp=sharing)

# **Introduction :**

The Heart Disease Prediction dataset is a collection of medical

records from patients who were evaluated for heart disease at the

Cleveland Clinic Foundation in the United States. The dataset

contains 303 samples with 13 features, including patient age, sex,

blood pressure, serum cholesterol levels, and more. The target

variable is a binary label indicating whether or not the patient has

heart disease.

# **Problem Statement :**

The problem statement for the Heart Disease Prediction dataset is to

predict whether a patient has heart disease based on several medical

attributes such as age, sex, blood pressure, cholesterol level, and

others. This is a classification problem, where the output variable is

binary (1 if the patient has heart disease, 0 if not). The goal is to build

a machine learning model that can accurately predict whether a

patient has heart disease or not, using the available medical attributes.

This model can potentially help doctors and healthcare professionals

make more informed decisions and improve patient outcomes.

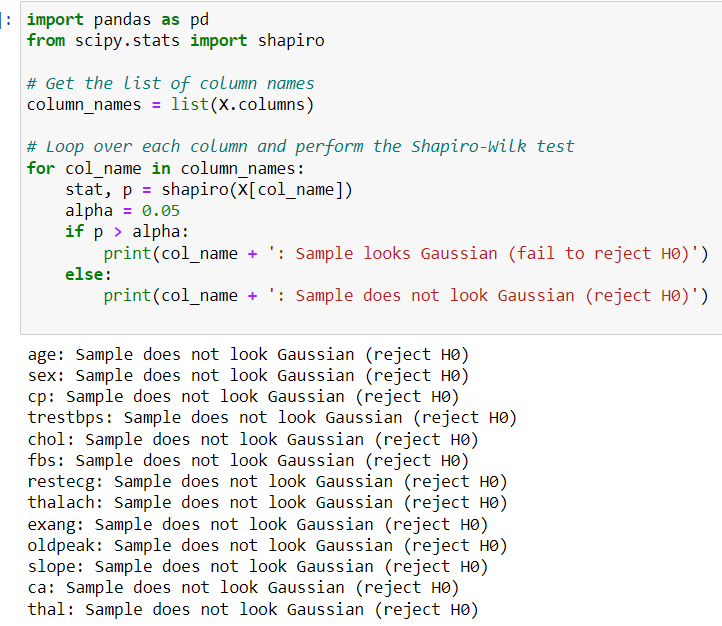
# **Objective :**

The objective of the heart disease prediction problem is to build a machine learning model that can accurately predict whether a person is suffering from heart disease or not based on a set of medical attributes such as age, sex, blood pressure, cholesterol levels, etc. The ultimate goal is to improve early detection and prevention of heart disease, a leading cause of death worldwide.

# **Methodology :**

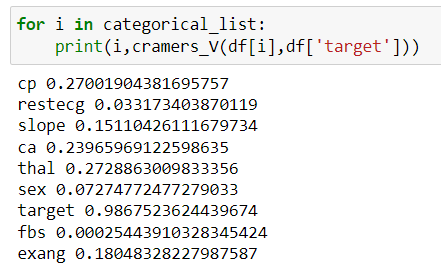
1. **Shapiro-Wilks Test:**

The Shapiro-Wilk test is a statistical test used to determine whether a given sample of data is drawn from a normal distribution. If the data is not normally distributed, these tests may not provide accurate results.



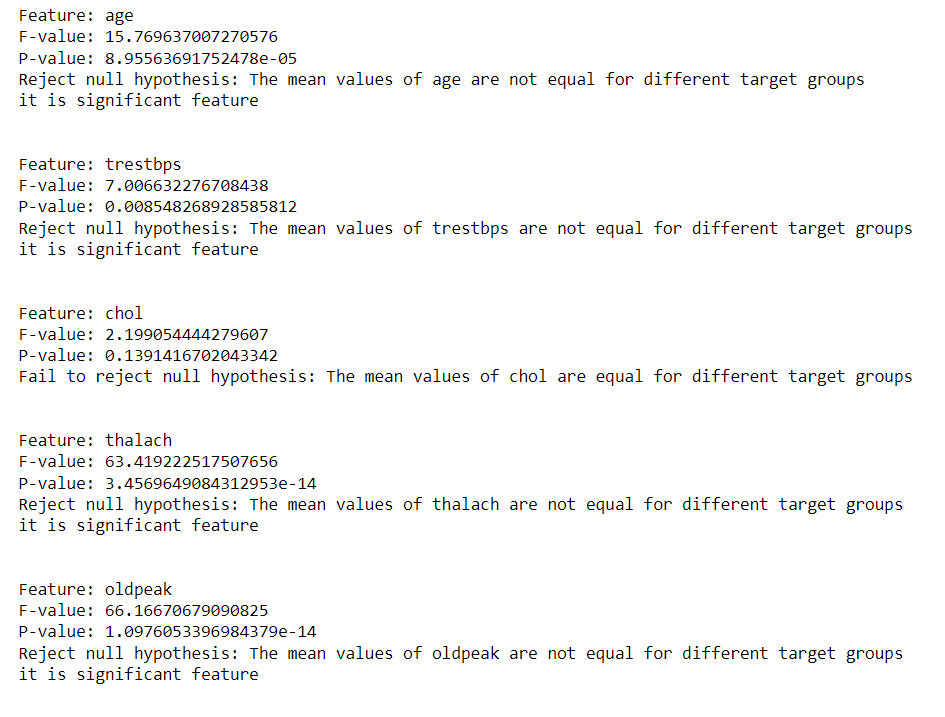
1. **Cramer - V Test:**

The test is particularly useful in cases where the variables have more than two categories and the chi-squared test cannot be used to measure the strength of association. The Cramer's V test helps to overcome the limitation of the chi-squared test by providing a measure of the effect size that can be interpreted in a similar way as the correlation coefficient.



1. **ANOVA Test:**

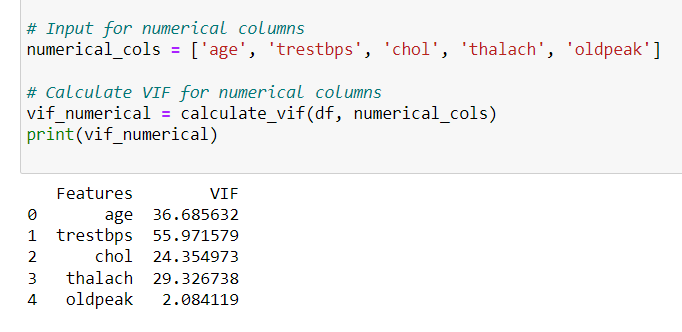
ANOVA (Analysis of Variance) test is a statistical method used to determine whether there are any significant differences between the means of two or more groups.

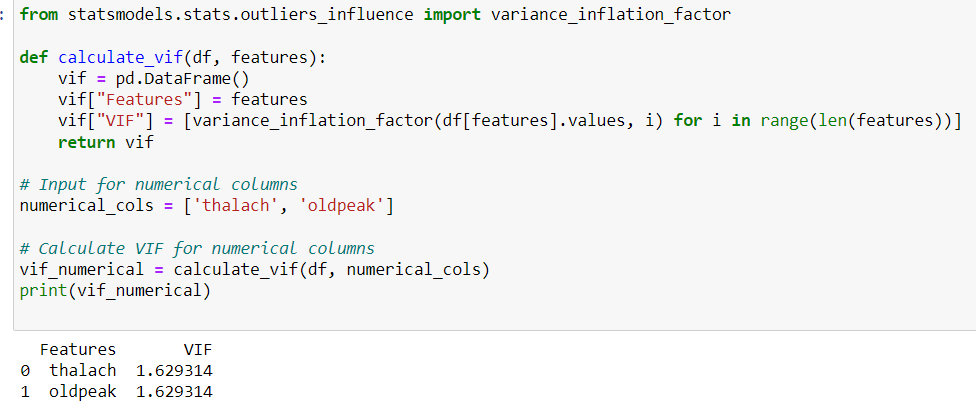


1. **Variance Inflation Factor (Multicollinearity Check):**

VIF measures the extent to which the variance of the estimated regression coefficient is increased due to multicollinearity in the data.

A VIF value of 5 or higher indicates that there is a high degree of multicollinearity in the data.

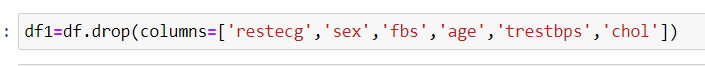




1. **Dummy Encoding :**

Dummy encoding, also known as one-hot encoding, is necessary for nominal variables with more than two classes because it allows us to represent each class as a separate binary feature.

# **Dropping insignificant features :**



# **Validation Used :**

# **Why K –fold instead of Train – test split :**

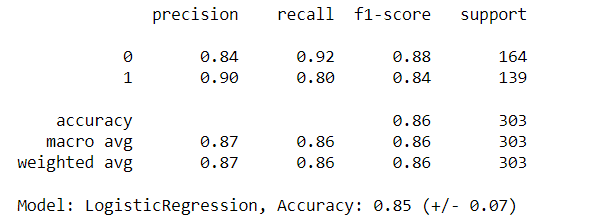
Train test split may not give an accurate estimate of the model's performance since it only evaluates the model on a single test set, which may not be representative of the overall data.

But,T he advantage of k-fold cross-validation is that it uses the entire dataset for training and testing, which can lead to a more accurate estimate of the model's performance.

## **Models and results before Optimization :**

1. **Logistic Regression:**

**Results :**

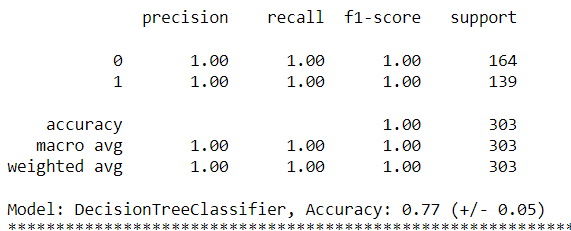


**Interpretation :**

The macro-average f1-score is 0.86, indicating that the model performs equally well for both classes. The weighted-average f1-score is also 0.86, indicating that the model is well-balanced in terms of correctly predicting both classes**.**

1. **Decision tree classifier :**

**Results :**

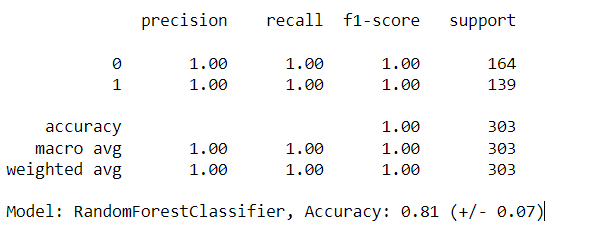


**Interpretation :**

It is possible that the DecisionTreeClassifier model is overfitting the training data.it is necessary to analyze other metrics and perform further testing to confirm if the model is indeed overfitting.

1. **Random Forest Classifier :**

**Results :**

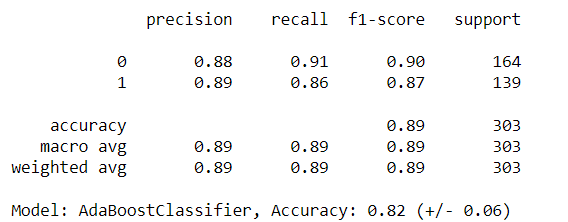


**Interpretation :**

It is possible that the Random forest classifier model is overfitting the training data.It is necessary to analyze other metrics and perform further testing to confirm if the model is indeed overfitting.

1. **Adaboost Classifier :**

**Results :**



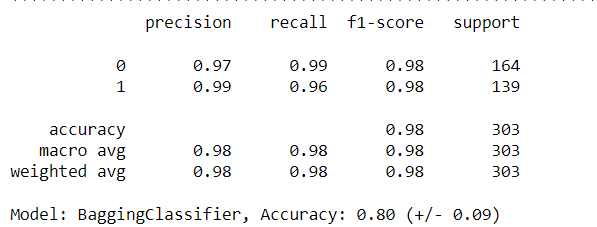
**Interpretation :**

AdaBoostClassifier model has a macro-average and weighted-average f1-score of 0.89, indicating a good balance between precision and recall for both classes.

Average accuracy of the model and the variance of the accuracy across different cross-validation folds. In this case, the model has an accuracy of 0.82 on average, with a standard deviation of 0.06 .

1. **Bagging Classifier :**

**Results :**

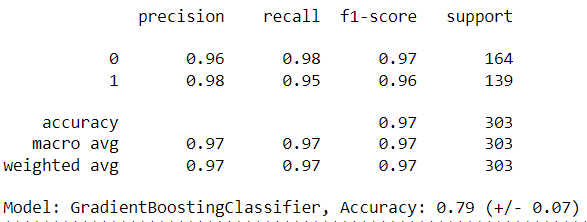


**Interpretation :**

It is possible that the DecisionTreeClassifier model is overfitting the training data, since it is achieving a perfect score (1.00) on both precision and recall for both classes, and also a perfect score on the overall accuracy. However, it would be necessary to analyze other metrics and perform further testing to confirm if the model is indeed overfitting.

1. **Gradient Boosting Classifier :**

**Results :**



**Interpretation :**

The model achieved an accuracy of 0.97 on the test data.

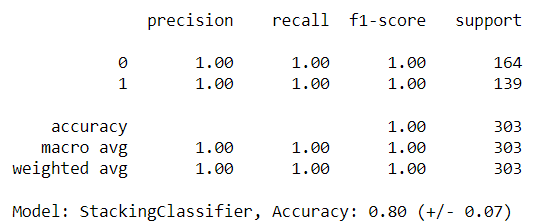
The precision for class 0 is 0.96, 96% were actually class 0.

The recall for class 0 is 0.98, which means that out of all the instances that were actually class 0, the model was able to correctly identify 98% of them.

The f1-score for class 0 is 0.97, which is the harmonic mean of precision and recall for that class.

1. **Stacking Classifier :**

**Results :**



**Interpretation :**

The classification report shows that the StackingClassifier model has a perfect accuracy of 1.0. It is necessary to analyze other metrics and perform further testing to confirm if the model is indeed overfitting.

## **Overall insights :**

Most of the models have high accuracy scores, with some variation in the precision, recall, and f1-score for each class.

It's worth noting that for All **tree based models** giving the accuracy score is very high at 100%, which may indicate that the models are **overfitting** the training data.

**Reason :**

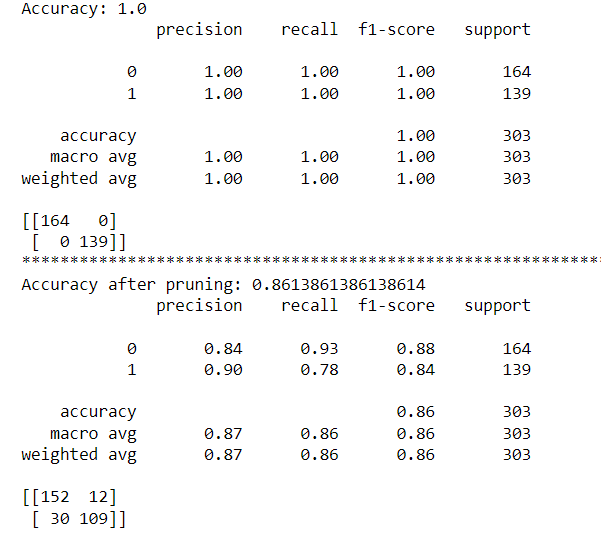
Tree-based models such as Decision Trees, Random Forests, and Gradient Boosting are prone to overfitting if the hyperparameters are not tuned properly.

## **Model Performance after Hyperparametric Tuning :**

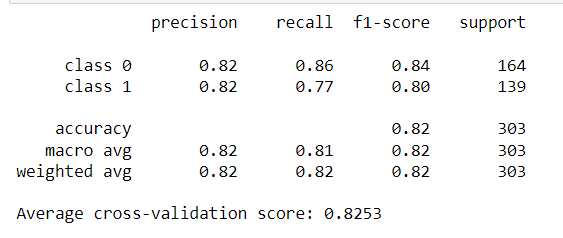
* After pruning, the decision tree becomes less complex, and the overfitting is reduced.
* The accuracy of the model was decreased slightly, but the model will be more generalizable and will perform better on new, unseen data.

**RESULTS AFTER PRUNING :**

Result of Random Forest Classifier :



Result of Stacking Classifier :



## **Best performed Model : Logistic Regression and why ? :**

* Logistic regression is a statistical method that is used to analyze the relationship between a binary outcome variable and one or more predictor variables.
* It is commonly used in classification tasks where the outcome variable is binary or categorical in nature.
* It is particularly useful when the relationship between the predictor variables and the outcome variable is not linear, but rather sigmoidal or S-shaped.
* In such cases, logistic regression is able to model the probability of the binary outcome as a function of the predictor variables, which makes it a powerful tool for classification tasks.

# **Applications**

# **Future scope of work**

Our work was limited products, finding expected number of banner ads to show the customer to make purchase. Further improvements can be done such as:

* Improve model performance
* Create loyalty programs
* Personalization
* Integration with other systems
* Recommendation
* Engage with customers on social media

# **Conclusion**

From the analysis, it appears that logistic regression performed the best in terms of accuracy and generalization to new data, followed by AdaBoost and Decision Tree Classifier ,Stacking classifier.