# Investigation of Future Cardiac Risk Prediction Using Machine Learning Based on SGPT Levels and Lifestyle Factors

**Devpal Singh Anand**
Amity University, Noida, India
*devpalsinghanand209@amity.edu.in*

***Abstract*--** The report that I am going to address everyone about, tends to include aspects of daily lifestyle habits, medical conditions it can lead to, incorporating the same into a machine learning model for prediction of Future Events of Cardiac Events based on an individual's levels of Serum Glutamic Pyruvic Transaminase a.k.a. SGPT. The creation of this report was inspired from insights of a life-changing podcast, consisting of valuable teachings from Dr. Shiv Kumar Sarin. Thus, leading to a creation of a simple but hard to form machine learning model, since the dataset was created completely from scratch. It included various key factors that influenced the levels of SGPT, being linked to liver diseases, that can lead to slight to even abnormal chances of cardiac events in an individual. The dataset designed for the study, consisted of generation of samples of 100 individuals using a custom logic incorporated in a python script taking help of the various key libraries such as Numpy, Pandas, Seaborn, and Matplotlib along-with several modules from Scikit-learn for incorporating various models and classifiers out of which DecisionTreeClassifier came on the top. In the latter phase, various evaluation metrics were taken into use for assessing the performance of the model, and the most accurate one was saved using Pickel for future use.

***Keywords:*** Cardiac events, machine learning model, SGPT, Daily lifestyle habits, Decision Tree Classifier



## I. Introduction

Frankly speaking, I searched everywhere for the dataset I required for analyzing the accuracy of my model but was not able to get my hands on any specific datasets that included the features with combination of SGPT / ALT enzyme levels that I needed for this particular research. Henceforth, I went on to google, visited various authentic medical sources, and even went through some social media platforms, in hope of gaining some beneficial insights. At first, I did not get much information, but as I delved deeper, I got more and more information.

### 1.1 Description of the research study

The aim of the research is mainly to bring the focus of all of us to a particular concern, that there is no need for any high-fidelity medicines or surgeries to get cured of diseases, if we just focus on getting our circadian rhythm into place, thus getting our lives on track.
Furthermore, all of you will be shocked to know that, if we just follow an overall healthy diet, then there is not even need for anything, we won't be needing any consulting too. Thus, my research mainly takes into consideration the relationship between the SGPT levels in an individual that have been affected due to hereditary illnesses or due to the unhealthy lifestyle followed by the individual. This does not mean, if you have got abs or good biceps, then you are healthy, not at all. Even if you are having a great body, and all, but are compromising other areas of the life you are living, then you are in great trouble, there are still chances of having cardiac events in a period of about 10 years counting from now. Thus, by observing the daily life habits of a person, we can estimate the SGPT levels, estimating if it can lead to liver diseases, and how much it can contribute to the level of cardiac events.

### 1.2 Problem statement

The problem I had to face here was that most of the datasets I had been working on, and even researching presently for this report, consisted mainly of the numerical values rather than the categorical ones, such as - considering the cholesterol levels, other medical terms and what not. But, I had not much experience in the health sector, so, had to make use of multi-class variables, which made it a bit harder for me. But, after extensive trial and error, creating various logical algorithms that match with the prediction model of mine, was a headache.

### 1.3. Motivation

The more and more I worked on this research of mine, it was getting interesting for me. But, as soon as I reached the part where there was a need to generate the logic for the model as mentioned earlier, it took away all the motivation, and confidence I had to such an extent that I deleted the whole 350 lines of logic I had created and slept. Then later on, I made use of our all-new Chat GPT 4o, and attached my dataset and asked why I am facing this issue. It provided me with a simple reason, that it could be due to oversampling or inequality between the samples provided, because of which the issue of accuracy is being faced.

As soon as I saw these statements, it struck me as a silly mistake, and as I corrected it, I eventually added more key features that were required and tweaked the logic a bit. To my surprise, it worked and boosted up my motivation further and led to a creation of an accurate but a different kind of working model. Since then, the algorithm has changed.

### 1.4 Key contributions

The key contributions in this report study are as follows:

1. Compilation of Data: - This includes compilation of data created with the help of logic generated from scratch.

2. Making use of a Machine Learning Model: - Here, generation of a prediction model is done, to predict the outcomes, if the individual will be having a future risk of cardiac event or not.

3. Implementation, Analysis and Validation of Code: - Implementation of algorithm, making use of classification reports, correlation analysis, and confusion matrix for the further validation of the model.

### 1.5 Organization of the paper

The paper is organized as follows: -

1. Introduction - This includes the overview of the report, the description of the field of research as well as the problems related to it, with a touch of motivation and key contributions linked to it.

2. Review of Literature - This includes the background of the whole idea of research, how it arose and conclusions it led to.

3. Materials and Methods - This refers to the methodologies proposed, algorithms and flowcharts that were drawn, etc.

4. Experimental Analysis - This section includes all the supervised classifiers that were used to generate analysis for the same.

5. Result discussion - This area focuses on the evaluation metrics and the visuals generated for deep insights of the dataset as well as provides the audience with a confidence that the report work is genuine.

## II. Literature review

The literature revolves around the theories, and research gathered from various authentic sources present in the studies of the authors, published for the people to see, and make further analysis upon.

## 2.1 Background

The background of the study takes into picture the relationship between the SGPT levels, and the severity of Liver Diseases, and the link of the severity of liver diseases to with that of the Future risks of cardiovascular events that can occur in an individual, which one cannot make out, just by observing the physical fitness of the person. Since, the outer body is nothing, but a reflection of what the individual wants the world to perceive of him/herself. Thus, when we talk about a person's health it depends on external as well as internal factors that influence the person's mental, physical, physiological, and psychological phases the individual goes through from various stages of life.

The inspiration for the creation of this report delved from a random encounter while strolling through 'YouTube' - a social media application, shorts, wherein there was a Podcast featuring Dr. Shiv Kumar Sarin, a renowned medical professional. During the podcast, the doctor shared his views on two particular aspects, namely the - 'Overall Functioning of the Liver', and 'Leading a long healthy life'.

As the interview continued, discussion of metabolic abnormalities came into picture, where he shared the simplest formula for an individual to check if he or she is an individual following an obese lifestyle or a healthier one. That is Height (in cms) - 100 for an individual suffering from any illness or lifestyle imbalance, and Height (in cms) - 105 for an individual suffering from any illness or lifestyle imbalance, in addition to the effects of hereditary illnesses that are over-powering the discipline the person is trying to bring into one's life.

In the latter phase, the doctor emphasized the significance of the SGPT test of just costing about 20-25 rupees, stating it as a one of the key indicators of liver dysfunction. He even highlighted a concerning scenario, wherein if a 26-yr old individual's SGPT level is 80, normal range being 30, there is a 7 times more chances of the person experiencing a cardiac event in about 10 years from now, i.e., near about 36-40 yrs of age, even if the person appears to be fit, and attends gym regularly. This particular insight intrigued me to delve deeper into the aspects of the linkage between the levels, and the future risks of the cardiovascular diseases, sparking an idea into me of researching more about the key factors & enzymes that can affect an individual's internal processes leading to risks of cardiac arrest, and gave me an idea of developing a machine learning model for prediction of the same.

Going further into the study, I discovered that it is only the SGPT levels that are to be considered, but there is also a need to reflect upon the interconnected nature of the other key enzymes too, such as the SGOT (Serum Glutamic Oxaloacetic Transaminase a.k.a. AST), and GGT a.k.a. Gamma Glutamyl Transferase. Deeply researching the enzymes, it came to my notice that, chronic liver conditions often lead to a systemic inflammation of the internal of the body as a whole (that cannot be observed externally) as well as metabolic imbalances, that in turn intensifies the risk of cardiovascular diseases. The studies of (Targher et al. ,2010), and

(Lonardo et al. ,2016) it is observed that individuals suffering from severe liver diseases have a higher chance of developing cardiac complications.

Still, the thought of a physically fit individual also having a 7 times chance of developing a future risk of cardiovascular disease was roaming in my mind. So, just to clear my doubts of everything and creating a proper simple-working prediction model, I used multiclass features, rather than numerical ones such as - Gender, Smoking habits, eating habits, drinking habits, Sleep Schedule, Medication effects, Severity of liver diseases & Age, and SGPT levels being the numerical ones, that have higher chances of affecting an individual's health contributing to future risk of cardiac events.

## III. Materials and methods

In this particular section, I have included all the materials and the various types of methodologies I had used, to gain insights from, for the further analysis in my study.

### 3.1 Proposed methodology

This subsection includes the detailed steps taken in achieving the objectives of the study, that include the process of data collection, preprocessing of data, model development, and the evaluation techniques.

#### 3.1.1 Data Collection

Here, a custom data set was generated for gaining a deeper understanding of the field of research, since there was no such research found, that created a link between the high levels of SGPT, predicting increased chances of future risks of cardiovascular diseases.

The unavailability of the specific links that I needed to use to fill the gaps between the passage of severity of liver diseases to future risks of cardiac arrest.

The dataset was created using python script, making use of libraries such as Numpy, Pandas, Seaborn, and Matplotlib.

#### 3.1.2 Data Preprocessing

This includes the preprocessing steps such as Cleaning of data, to check for outliers, handling missing values, and normalizing the numerical features; Feature Engineering as well as Splitting the dataset into training and testing for further analysis for testing the accuracy, and reliability of the predictions.

1. **Cleaning the dataset -**

   Here, the cleaning was not required, since the dataset was generated from scratch. Henceforth, had used such a logic that does not give me any missing or null values, that could deter the prediction.

2. **Feature Engineering -**

   The above step was done, but it does not indicate, the dataset was up to the mark, it was imbalanced, i.e., the outcomes of the future risks were imbalanced to such an extent that, even when the SGPT levels were low, the future risks of cardiac events were shown as high. This created a need to add more features and modify the script totally. I even had to tweak the logic used such that, it included the multiclass variables, that included the lifestyle habits of an individual such as (Eating habits, drinking habits, Exercise habits, sleeping habits, Smoking habits), and first I had made the impact of medications such as it affects the SGPT levels to rise. But then I modified the logic to be such that it affected the severity of liver diseases, lowering the SGPT levels, and providing the future risks of cardiac events based on the interpretations of the SGPT levels.

   The dataset had become reliable, and accurate thereafter, but still this is a beginner level data set, and there will be need to add numerical features to make the dataset more varied according to the various illnesses faced by the individuals present out there.

3. **Splitting the dataset -**

   I splitted the same dataset into two parts, first half as training dataset and the other half as test dataset, to check the accuracy of the prediction. Since, if the dataset is accurate enough, then even if any professional uses it to add the numerical features, making it more advanced, she or he won't face any issues with respect to the accuracy of the prediction, instead it will help the audience to get deeper insights on the dataset itself.

### 3.1.3 Exploring the Data

Here, I followed 3 major steps-

1. **Creation of Visualizations**
   Here, creation of visualizations such as pie-charts and bar graphs were done.

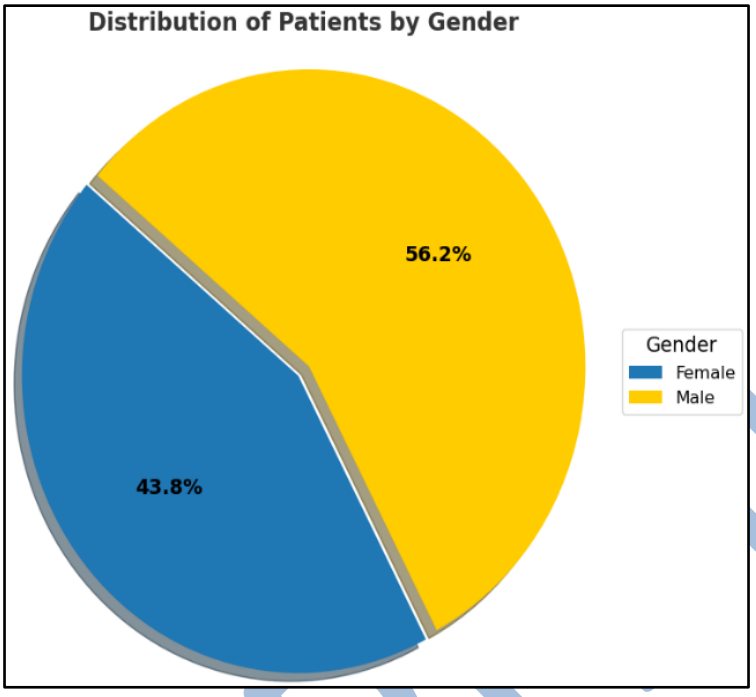



Figure 1: A pie chart showing distribution of patients by gender.

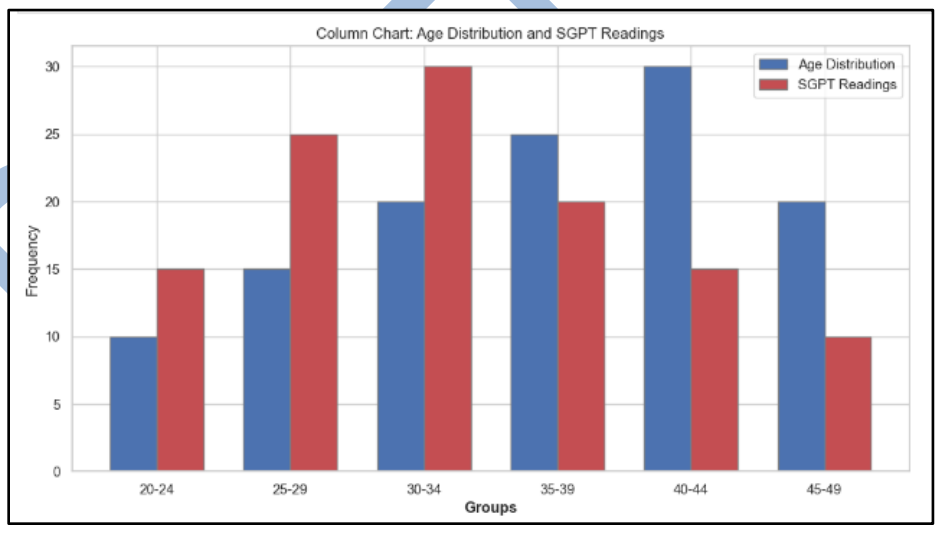



Figure 2: A column chart visually representing Age Distribution & SGPT Readings.

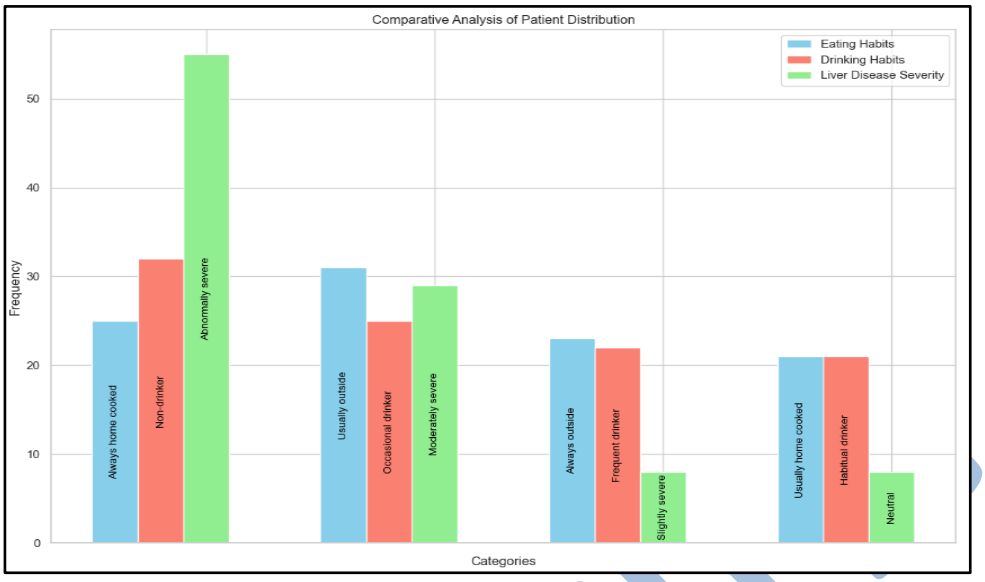



Figure 3: Comparative analysis of patient distribution by eating habits, drinking habits & liver disease severity.

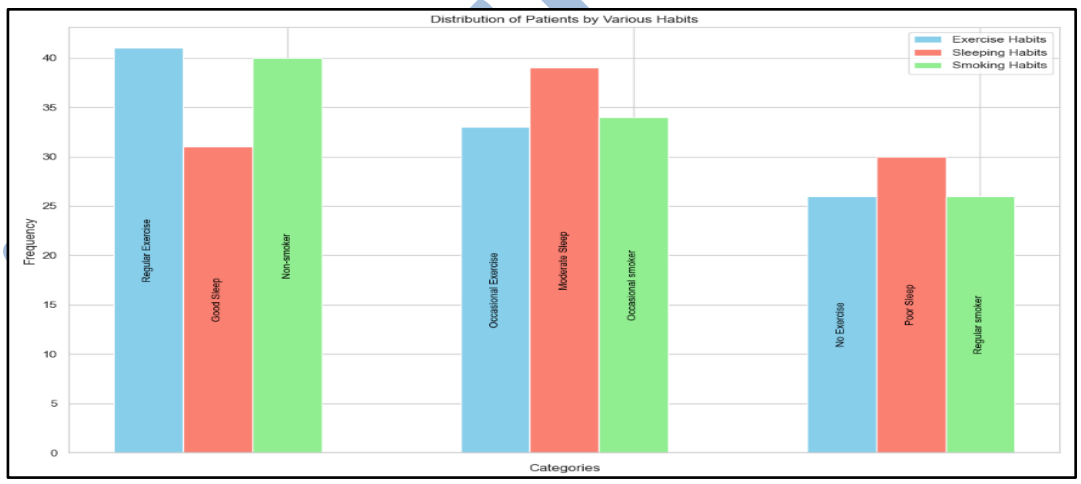



Figure 4: Comparative analysis of patient distribution by exercising habits, sleeping habits, and smoking habits.

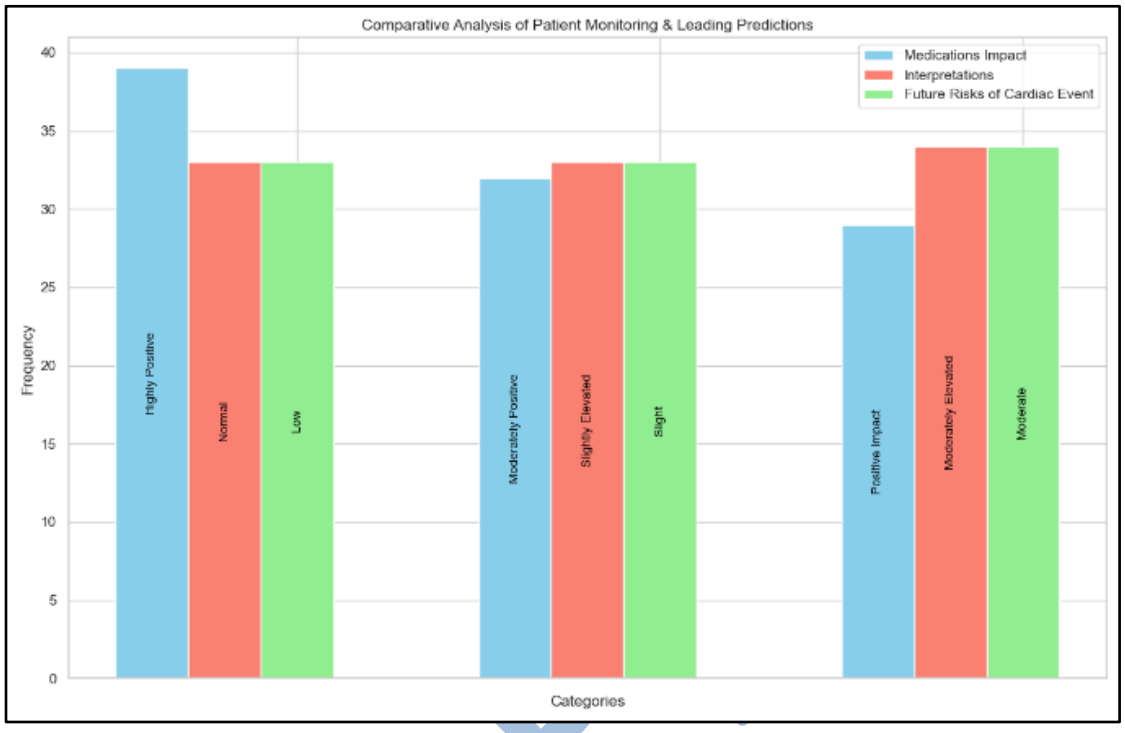



Figure 5: Comparative analysis of patient monitoring & leading predictions based on impact of medications, and interpretations of the impacts leading to future risks levels of the cardiac event.

## 2. Model Selection & tuning

Here, multiple machine learning models were tested thoroughly such as "Logistic Regression", "Kneighbors Classifier", "Support Vector Classifier", Gaussian NB", "Decision Tree Classifier", and "Random Forest Classifier", so that the predictions can prove out to be as accurate as possible.

In the end, the 'Decision Tree Classifier ()' model turned out to be the most accurate one.

Since, the model proved out to be accurate in the first attempt itself, tuning was not brought into picture, but, still to clear the doubts, as to how the data from scratch can be accurate, "AUC-ROC Curve" along-with "Confusion Matrix" were plotted to confirm the accuracy, whether it was due to the model being biased towards a particular class or it was evenly checking all the multi-class variables, making it accurate.

We will talk about each of the models deeply, along with their use in the "Experimental Analysis" section.

### 3.2 Algorithm

**=> [Add the step-by-step algorithm and describe it]. [Tmr] [ Do from blackbox wtsp link]**

Here the terms of algorithm is as follows: -

1. Input Values: -

   **"diet":** refers to the diet preferred by any individual (Always home cooked, usually home cooked, Usually Outside, and Always Outside).

   **"dr":** refers to the drinking habits of any individual (Non-drinker, Occasional drinker, Frequent drinker, and Habitual drinker)

   **"exer":** refers to exercise routine followed by any individual (Regular Exercise, Occasional Exercise, No Exercise)

   **"sl":** referring to sleeping schedule followed by any individual (Good Sleep, Moderate Sleep, Poor Sleep)

   **"smok_hab":** refers to a bad habit of smoking pursued and abused by any individual (Non-smoker, Occasional smoker, Regular smoker)

   **"meds":** referring to the impact of medications given to the required individual on liver disease (Positive Impact, Moderately Positive, Highly Positive)

2. Output: -

   **"liver_disease":** severity of liver disease (Neutral, Slightly severe, Moderately severe, Abnormally severe)

   **"sgpt":** refers to SGPT levels in numerical form.

   **"future_risk":** refers to the future risk of cardiac events predicted based on the levels of SGPT (Low, Slight, Moderate, High)

   **"interpretation":** refers to the interpretation of future risk of cardiac events based on its predicted values.

   **"notes":** refers to the notes for monitoring the cardiovascular health of any individual (Monitor Yearly, Monitor Monthly, Monitor Weekly, Monitor Daily)

Algorithm is as follows:

I. Determining Severity of Liver Disease: -

```
if diet == 'Always home cooked' and dr == 'Non-drinker':
    if exer == 'Regular Exercise':
        if sl in ['Good Sleep', 'Moderate Sleep']:
            liver_disease = 'Neutral'
        else:
            liver_disease = 'Slightly severe'
    elif exer == 'Occasional Exercise':
        if sl == 'Good Sleep':
```

```
            liver_disease = 'Neutral'
        else:
            liver_disease = 'Slightly severe'
    else:
        if sl == 'Good Sleep':
            liver_disease = 'Neutral'
        elif sl == 'Moderate Sleep':
            liver_disease = 'Slightly severe'
        else:
            liver_disease = 'Moderately severe'
elif diet == 'Always home cooked' and dr == 'Occasional drinker':
    if exer == 'Regular Exercise':
        if sl in ['Good Sleep', 'Moderate Sleep']:
            liver_disease = 'Neutral'
        else:
            liver_disease = 'Slightly severe'
    elif exer == 'Occasional Exercise':
        if sl == 'Good Sleep':
            liver_disease = 'Neutral'
        elif sl == 'Moderate Sleep':
            liver_disease = 'Slightly severe'
        else:
            liver_disease = 'Moderately severe'
    else:
        if sl == 'Good Sleep':
            liver_disease = 'Slightly severe'
        else:
            liver_disease = 'Moderately severe'
elif diet == 'Always home cooked' and dr in ['Frequent drinker', 'Habitual drinker']:
    if exer == 'Regular Exercise':
        if sl == 'Good Sleep':
            liver_disease = 'Slightly severe'
        else:
            liver_disease = 'Moderately severe'
    elif exer == 'Occasional Exercise':
        if sl == 'Good Sleep':
            liver_disease = 'Moderately severe'
        else:
            liver_disease = 'Abnormally severe'
    else:
        if sl == 'Good Sleep':
            liver_disease = 'Moderately severe'
        else:
            liver_disease = 'Abnormally severe'
elif diet == 'Usually home cooked':
    if dr == 'Non-drinker':
```

```
        if exer == 'Regular Exercise':
            if sl in ['Good Sleep', 'Moderate Sleep']:
                liver_disease = 'Neutral'
            else:
                liver_disease = 'Slightly severe'
        elif exer == 'Occasional Exercise':
            if sl == 'Good Sleep':
                liver_disease = 'Neutral'
            else:
                liver_disease = 'Slightly severe'
        else:
            if sl == 'Good Sleep':
                liver_disease = 'Slightly severe'
            else:
                liver_disease = 'Moderately severe'
    elif dr == 'Occasional drinker':
        if exer == 'Regular Exercise':
            if sl == 'Good Sleep':
                liver_disease = 'Neutral'
            else:
                liver_disease = 'Slightly severe'
        elif exer == 'Occasional Exercise':
            if sl == 'Good Sleep':
                liver_disease = 'Slightly severe'
            else:
                liver_disease = 'Moderately severe'
        else:
            if sl == 'Good Sleep':
                liver_disease = 'Moderately severe'
            else:
                liver_disease = 'Abnormally severe'
    else:
        if exer == 'Regular Exercise':
            if sl == 'Good Sleep':
                liver_disease = 'Moderately severe'
            else:
                liver_disease = 'Abnormally severe'
        elif exer == 'Occasional Exercise':
            liver_disease = 'Abnormally severe'
        else:
            liver_disease = 'Abnormally severe'
elif diet == 'Usually outside':
    if dr == 'Non-drinker':
        if exer == 'Regular Exercise':
            if sl == 'Good Sleep':
                liver_disease = 'Slightly severe'
```

```
            else:
                liver_disease = 'Moderately severe'
        elif exer == 'Occasional Exercise':
            liver_disease = 'Moderately severe'
        else:
            liver_disease = 'Abnormally severe'
    elif dr == 'Occasional drinker':
        if exer == 'Regular Exercise':
            liver_disease = 'Moderately severe'
        else:
            liver_disease = 'Abnormally severe'
    else:
        liver_disease = 'Abnormally severe'
elif diet == 'Always outside':
    if dr == 'Non-drinker':
        if exer == 'Regular Exercise':
            if sl == 'Good Sleep':
                liver_disease = 'Slightly severe'
            else:
                liver_disease = 'Moderately severe'
        elif exer == 'Occasional Exercise':
            liver_disease = 'Moderately severe'
        else:
            liver_disease = 'Abnormally severe'
    elif dr == 'Occasional drinker':
        liver_disease = 'Abnormally severe'
    else:
        liver_disease = 'Abnormally severe'
```

II. Determining Liver Disease Severity based on Smoking habits: -

```
if smok_hab == 'Non-smoker':
    pass
elif smok_hab == 'Occasional smoker':
    if liver_disease == 'Slightly severe':
        liver_disease = 'Moderately severe'
elif smok_hab == 'Regular smoker':
    if liver_disease == 'Slightly severe':
        liver_disease = 'Abnormally severe'
    elif liver_disease == 'Moderately severe':
        liver_disease = 'Abnormally severe'
```

III. Determining levels of SGPT on basis of Severity of Liver Diseases: -

```
if liver_disease == 'Slightly severe':
    sgpt += 7
elif liver_disease == 'Moderately severe':
    sgpt += 15
elif liver_disease == 'Abnormally severe':
    sgpt += 25
```

IV. Determining levels of SGPT on basis of impact of medications: -

```
if meds == 'Positive Impact':
    sgpt -= 5
elif meds == 'Moderately Positive':
    sgpt -= 10
elif meds == 'Highly Positive':
    sgpt -= 15
```

V. Determining future risk of cardiac events on the basis of levels of SGPT: -

```
if sgpt < 40:
    future_risk = 'Low'
elif sgpt >= 40 and sgpt < 60:
    future_risk = 'Slight'
elif sgpt >= 60 and sgpt < 80:
    future_risk = 'Moderate'
else:
    future_risk = 'High'
```

VI. Determining interpretation of the level of cardiac event based on the future risks that were predicted earlier: -

```
if future_risk == 'Low':
    interpretation = 'Normal'
elif future_risk == 'Slight':
    interpretation = 'Slightly Elevated'
elif future_risk == 'Moderate':
    interpretation = 'Moderately Elevated'
else:
    interpretation = 'Abnormal'
```

VII. Creating notes on basis of interpretations analyzed from the future risk: -

```
if interpretation == 'Normal':
    notes = 'Monitor Yearly'
elif interpretation == 'Slightly Elevated':
    notes = 'Monitor Monthly'
elif interpretation == 'Moderately Elevated':
    notes = 'Monitor Weekly'
else:
    notes = 'Monitor Daily'
```

**3.3 Dataset description**

Here, the dataset named health_consistent.csv, consists of a total of 13 variables out of which 10 are the "Feature Variables" on the basis of which the one (1) "Target Variable" is to be predicted.

The other 2 variables that are left out or dropped are the "Notes" & "Interpretation" variables, these are not counted as the feature ones, since these act as the "Side Variables", and could have created a discrepancy while testing out the prediction model.

So, now we were left with:
A, Feature Variables: "Gender", "Age", "Eating Habits", "Drinking Habits", "Exercise Habits", "Sleeping Habits", "Smoking Habits", "Liver Disease Severity", "Medication Impact", "SGPT Readings".
B. Target Variables: "Future Risks of Cardiac Event"

Here, the "Age" and "SGPT Readings" are in numerical ones, and the other features as mentioned above are multiclass variables, being categorical.

The "Eating Habits". "Drinking Habits", and the "Liver Diseases" are divided into 4 sub-categories.

The "Exercise Habits", "Sleeping Habits", and the "Smoking Habits" are divided into 3 sub-categories.

In addition to this, all the features are interlinked to each other, as in, the "Gender & Age" are have some effect on the liver being old as well as the levels of SGPT, since, it is found in a research **[3]** that levels of SGPT is high in males as compared to females, and age also has a significant role in it.

If we go further in the case of the habits, the "Eating, Drinking, Exercising, Sleeping, Smoking", these habits have a great effect on one's liver as well as the circadian rhythm that is responsible for keeping one's body in balance, and if it is not taken care of it can affect the liver starting from normal to severe levels, based on which the SGPT levels will have no effect in case of a disciplined individual, but, will have great impact in case of a individual following a unhealthy lifestyle.

Later, comes the impacts of the medication, the impact can be either positive or more positive, since, if the impact is negative, the professionals handling the individual will change the medications as soon as they notice if it is having an opposite effect or no impact on the individual's liver health at all, thereby, lowering the individual's levels of SGPT.

Then, based on the readings of SGPT, the Future risks of cardiac events will be predicted, that will be 10 years from the time of prediction, and interpretations will be made, and the individual will be advised as to how frequently she/he has to monitor their levels of SGPT.

## IV. Experimental Analysis

=> Refers to the models and classifiers one uses for the prediction of their model.

### 4.1 Glimpse of the Models and Classifiers used -

#### 4.1.1 Logistic Regression: -

When we talk about logistic regression **[4]**, that is also known as the logit, MaxEnt classifier that is responsible for predicting the probability or chances of an outcome or occurring of an event or observation, based on the binary or mathematical classification tasks being taken place to find relationships between 2 or sometimes more (multi-class) variables, being categorical.

#### 4.1.2 K Neighbors Classifier: -

Here, when I talk about the KNeighborsClassifier, it comes to my notice that it is nothing but a test that recognizes a pattern and carries out classification **[5]**, based on all its neighboring classifications, storing the available cases. Thus, even after mostly being recommended for Classification problems, it is also used in some

regression classifications too. One thing that makes it unique is that it does not make any assumptions, being a "non-parametric" algorithm.

### 4.1.3 Support Vector Classifier: -

As the trend grows, the support vector machine **[6]** has been introduced in the world of machine learning, that is based on the statistical theory of learning. It might not work as good as the Decision Tree Classifier, but it is proven to provide high accuracy and hence, have reached a good amount of global promotions.

### 4.1.4 Gaussian NB: -

This classifier officially being a part of the Naïve Bayes follows the normal distribution, while supporting the continuous data that flows along-with it **[5]**. Thus, the classifier has become the most common, since, it is the simplest of all the others in case of implementation as here the user is required to calculate just the statistical means of mean and standard deviations for data that was split for training data, thereby, helping in eliminating the insignificant specifications, thus providing better overall performance.

### 4.1.5 Decision Tree Classifier: -

When we talk about the DecisionTree as a visual representation **[5]**, we see that we are getting all the possible solutions to our problems of the machine learning models based on the given features we have provided. Henceforth, making it a multi-use model that can be used for both regression and classification analysis of the specific model it is being used. Some people may be curious as to why the name Decision "Tree", tree comes into picture, since it comprises two nodes, a decision node and Leaf node, where both have their own responsibilities to be taken care of respectively.

### 4.1.6 Random Forest Classifier: -

The Random Forest Classifier came into picture when the Decision Tree one was providing accuracy based on generalizations, **[7]** that could be either due to imbalance, overfitting or the samples being just perfect, which was hard to make out. Henceforth, Ho in the year 1995 introduced a random-subspace method, presented later as "Random Forest" by Breiman in 2001 in his studies. The advantage of this over "Decision Tree Classifier" is that this classifier focuses on many individual trees rather than one, which is also known as bootstrap aggregating or bagging, reducing the case of overfitting.

### 4.2 Tabular Representation

| **Classifier** | **Approach (Binary)** | **Approach (Multiclass)** | **Main Parameters** |
|---|---|---|---|
| LogisticRegression Classifier | Supervised learning | OneVsRest (OvR) or Softmax for multiclass | 'C', 'solver', 'max_iter' & 'penalty' |
| KNeighbors Classifier | Instance - based learning | Instance - based learning | 'n_neighbors', 'weights', 'algorithm' & 'p' |
| Support Vector Classifier (SVC) | Supervised learning (SVM) | OneVsRest (OvR) or OneVsOne (OvO) | 'C', 'kernel', 'gamma' & 'degree' |
| Gaussian Naive Bayes Classifier | Probabilistic learning (NB) | Probabilistic learning (NB) | 'var_smoothing' |
| DecisionTree Classifier | Supervised learning | Supervised learning | 'criterion', 'max_depth', 'min_samples_split' & 'min_samples_leaf' |
| RandomForest Classifier | Ensemble learning | Ensemble learning | 'n_estimators', 'criterion', 'max_depth', 'min_samples_split' & 'min_samples_leaf' |



## V. Result discussion

=> In this particular section the Dataset is split into the ratios of 80 by 20 for training and testing data respectively, which is then processed further such that the model gets fit with the training data, to make sure that the evaluation of the test data is not random or there are no chances for a doubt about the accuracy of the prediction.

The further analysis and validation is processed by carrying out various performance metrics as a whole using a classification report, that is mainly derived from the confusion matrix that helps in knowing that the model is legitimate and there is no sign of doubt.

### 5.1 Confusion Matrix

=> The confusion matrix being the most reliable & effective tool in analyzing predictions, is deployed to perceive the varied behavior of the different classifiers.

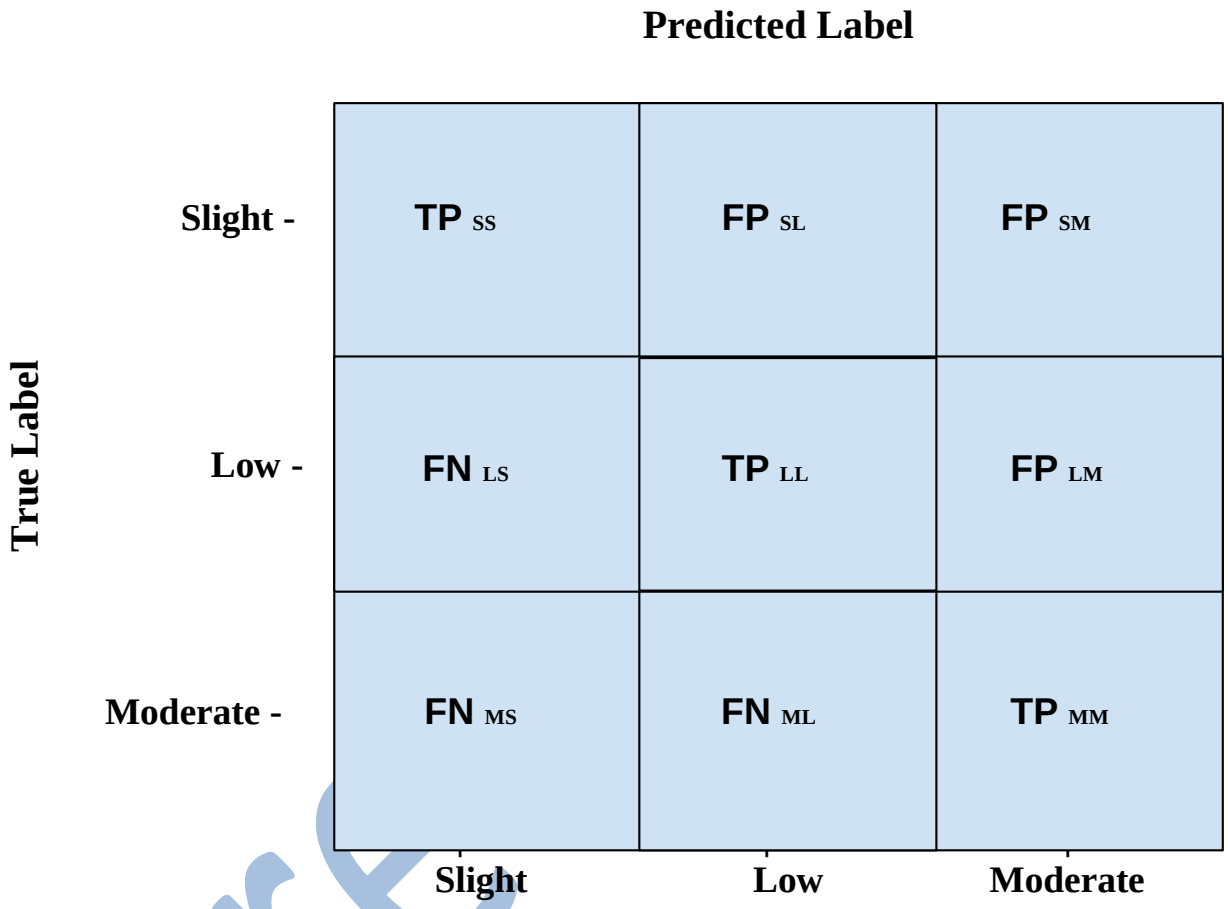



Figure 6: A Sample for 3 by 3 multiclass Confusion Matrix

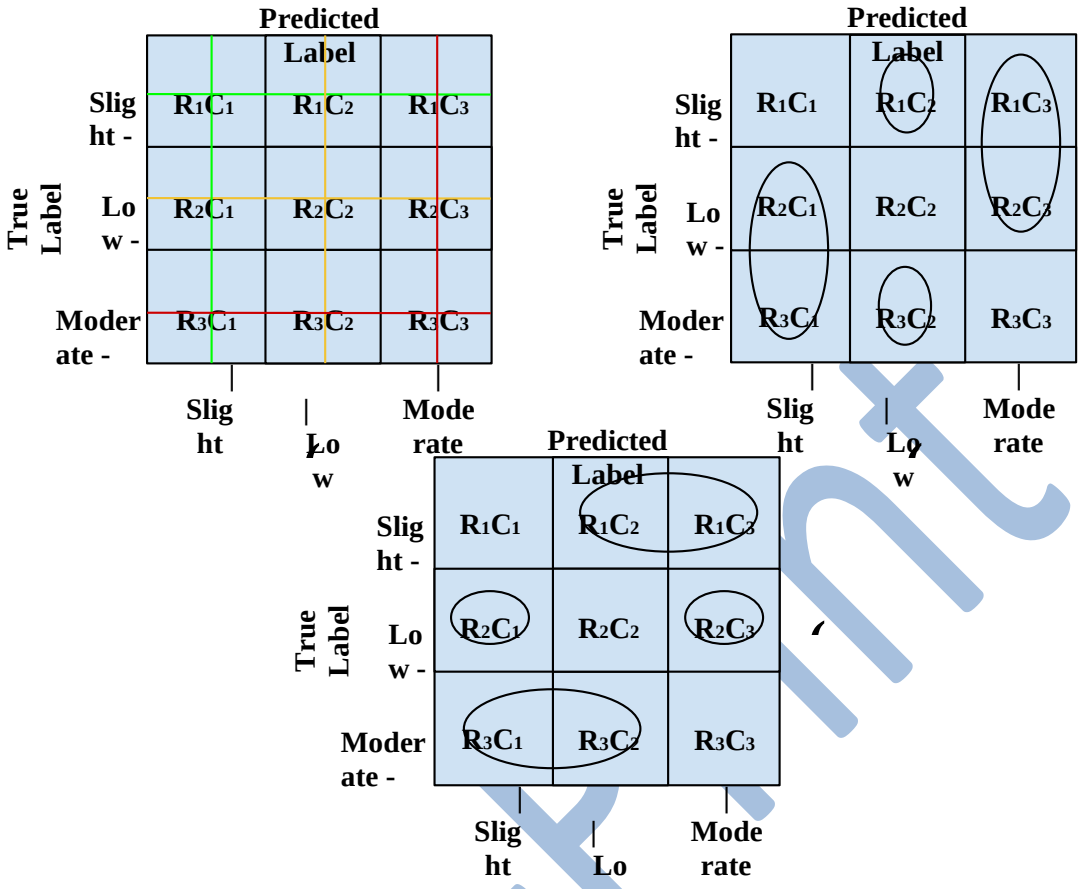



Figure 7: (A), (B) & (C) indicate the "TN", "FP" & "FN" values respectively.

Here, Fig. 7 depicts the calculation of the values for (A) - "True Negatives", (B) - "False Positives" and (C) - "False Negatives" values respectively. Since, the middle diagonals always tend to be termed as "True Positives", it is never a headache to calculate it out the hard way. The Calculation for each of it are as follows: -

1. **True Positives for each class: -**

=> Slight = $R_1C_1$; Low = $R_2C_2$ **;** Moderate = $R_3C_3$

Combined TP = $R_1C_1 + R_2C_2 + R_3C_3$

2. **True Negatives for each class: -**

=> Slight = $R_2C_2 + R_2C_3 + R_3C_2 + R_3C_3$

Low = $R_1C_1 + R_1C_3 + R_3C_1 + R_3C_3$

Moderate = $R_1C_1 + R_1C_2 + R_2C_1 + R_3C_2$

Combined TN = TN (Slight) + TN (Low) + TN (Moderate)

3. **False Positives for each class: - (**excludes the actual value & adds up the respective column values**)**

=> Slight = $R_2C_1 + R_3C_1$

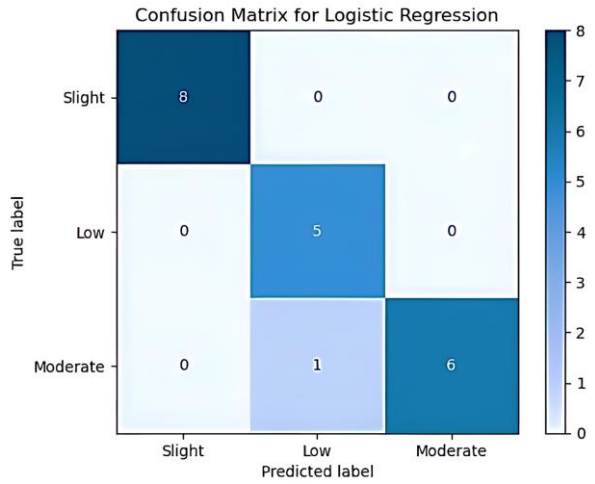
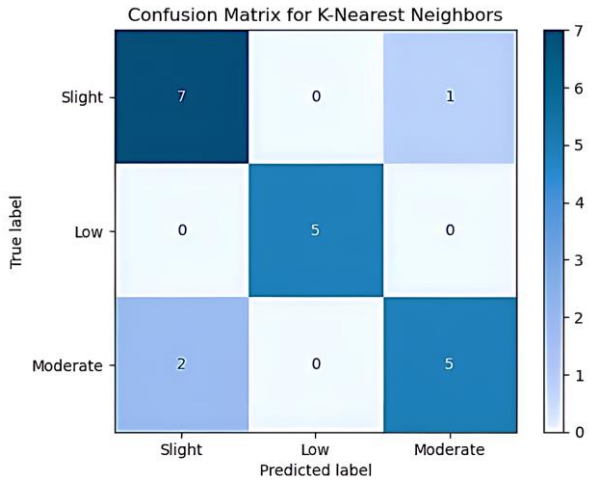
Low = $R_1C_2 + R_3C_2$

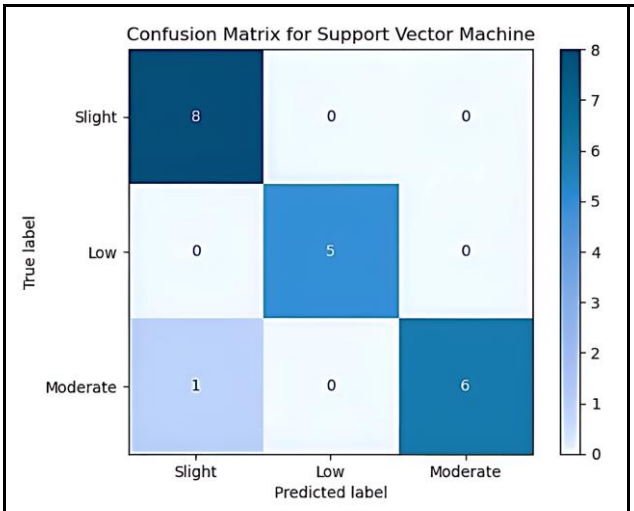
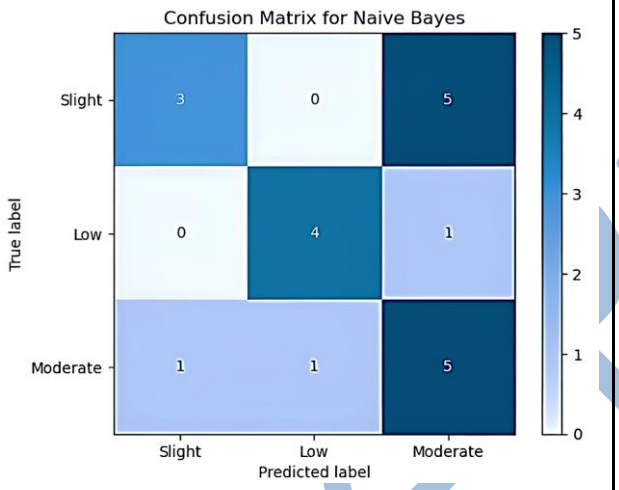
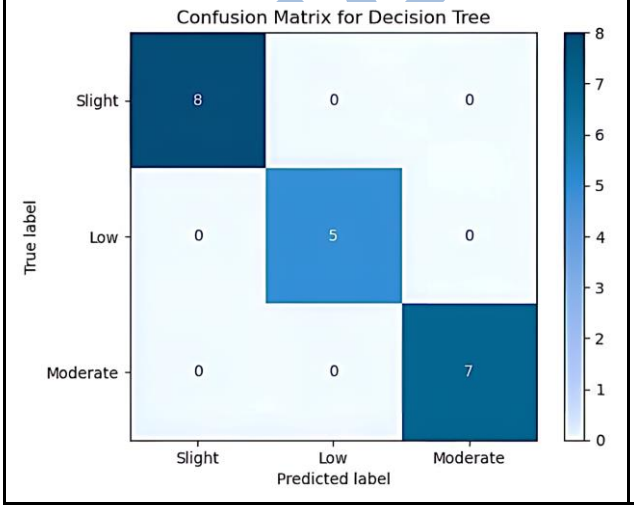
Moderate = $R_1C_3 + R_2C_3$
Combined FP = FP (Slight) + FP (Low) + FP (Moderate)

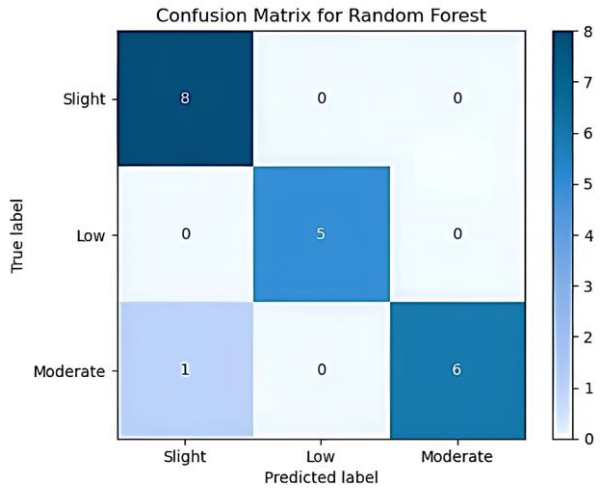
**4. False Negatives for each class: - (**excludes the actual values & adds up respective row values**)**

=> Slight = $R_1C_2 + R_1C_3$
Low = $R_2C_1 + R_2C_3$
Moderate = $R_3C_1 + R_3C_2$
Combined FN = FN (Slight) + FN (Low) + FN (Moderate)

Table 1. Description of Each of the Respective Model's Confusion Matrix

| Confusion Matrix | Description |
|---|---|
|  | **Total No. of Records** = 20<br><br>**TP (Slight, Low, Moderate)** = 8, 5, 6<br>**TN (Slight, Low, Moderate)** = 12, 14, 13<br>**FP (Slight, Low, Moderate)** = 0, 1, 0<br>**FN (Slight, Low, Moderate)** = 0, 0, 1<br><br>For the "Slight" class the model performed perfectly, with no "False Positive or False Negative" value. However, in case of "Low", the model made a single error, predicting low for an instance that was not "Low", but, actually a "Moderate" one. In the case of "Moderate", the model made an error by failing to identify an instance as moderate and misclassifying it as another class (Low). |
|  | **Total No. of Records** = 20<br><br>**TP (Slight, Low, Moderate)** = 7, 5, 5<br>**TN (Slight, Low, Moderate)** = 10, 15, 12<br>**FP (Slight, Low, Moderate)** = 2, 0 , 1<br>**FN (Slight, Low, Moderate)** = 1, 0, 2<br><br>For the "Slight" class the model has minor confusion, with few misclassifications of moderate ones as slight and one instance of slight being misclassified. In the case of "Low", the model performed well. However, in the case of "Moderate" class, a few moderate ones were misclassified as slight and one instance of slight was incorrectly classified as moderate. |

| | |
|---|---|
|  | **Total No. of Records Records** = 20<br><br>**TP (Slight, Low, Moderate)** = 8, 5, 6<br>**TN (Slight, Low, Moderate)** = 11, 15, 13<br>**FP (Slight, Low, Moderate)** = 1, 0, 0<br>**FN (Slight, Low, Moderate)** = 0, 0, 1<br><br>For the "Slight" class, the model correctly identifies the 8 instances as slight, making one error by incorrectly classifying as moderate. However, the model performs flawlessly for the "Low" class. Coming to the "Moderate" class, the model classifies 6 instances as moderate correctly, but misclassifies one instance as slight. |
|  | **Total No. of Records Records** = 20<br><br>**TP (Slight, Low, Moderate)** = 3, 4, 5<br>**TN (Slight, Low, Moderate)** = 11, 14, 7<br>**FP (Slight, Low, Moderate)** = 1, 1, 6<br>**FN (Slight, Low, Moderate)** = 5, 1, 2<br><br>Here, we can observe that, model has significant difficulties with "Slight" and "Moderate" classes wherein, it misclassified 5 "Slight" instances as moderate as well as 6 instances (5 of "Slight" and 1 of "Low") as moderate ones. However, in the case of "Low" it can be seen that the model only makes error in each FP & FN, by misclassifying 1 moderate instance as low and 1 low instance as being moderate, performing reasonably for "Low" class |
|  | **Total No. of Records Records** = 20<br><br>**TP (Slight, Low, Moderate)** = 8, 5, 7<br>**TN (Slight, Low, Moderate)** = 12, 15, 13<br>**FP (Slight, Low, Moderate)** = 0, 0, 0<br>**FN (Slight, Low, Moderate)** = 0, 0, 0<br><br>This is the best model I have come across, wherein, the model perfectly identifies all the instances of "Slight", "Low" and "Moderate" with no misclassifications, thereby, providing a 100% accuracy. Also, it shows that there are no "False Positives" or "False Negatives" in the data, making it highly effective & reliable in working with the prediction. |




**Total No. of Records Records** = 20

**TP (Slight, Low, Moderate)** = 8, 5, 6
**TN (Slight, Low, Moderate)** = 11, 15, 13
**FP (Slight, Low, Moderate)** = 1, 0, 0
**FN (Slight, Low, Moderate)** = 0, 0, 1

For the "Slight" class, the model correctly identifies the 8 instances as slight, making one error by incorrectly classifying as moderate. However, the model performs flawlessly for the "Low" class. Coming to the "Moderate" class, the model classifies 6 instances as moderate correctly, but misclassifies one instance as slight.

## 5.2. Performance Metrics

### 5.2.1. Accuracy

=> It measures how well the model performs overall by correctly identifying both positive and negative cases.

=> Formula:- $\frac{TP + TN}{TP + TN + FP + FN}$

### 5.2.2. Specificity

=> It refers to the proportion of the true negatives correctly identified out of all the negative instances.

=> Formula:- $\frac{TN}{TN + FP}$

### 5.2.3. Sensitivity / Recall

=> It measures as to how well the model identifies the positive cases.

=> Formula:- $\frac{TP}{TP + FN}$

### 5.2.4. Precision

=> It measures the overall accuracy of the positive predictions made by the respective model, depicting how many of the predictive outcomes were actually positive.

=> Formula:- $\frac{TP}{TP + FP}$

### 5.2.5. F1-Score

=> It refers to the calculation of harmonic mean of the Precision and Recall, balancing both the metrics, providing the user with a single metric that considers both the false positives and the false negatives.

=> Formula:- $2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$

| Classification Model | Accuracy | Specificity | Sensitivity / Recall | Precision | F1-Score | Error Rate |
|---|---|---|---|---|---|---|
| **Logistic Regression Model** | **Slight:** 1.00<br>**Low:** 0.95<br>**Moderate:** 0.95 | **Slight:** 1.00<br>**Low:** 0.93<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 0.86 | **Slight:** 1.00<br>**Low:** 0.83<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 0.91<br>**Moderate:** 0.92 | **Slight:** 0.00<br>**Low:** 0.05<br>**Moderate:** 0.05 |
| **KNeighbors Classification Model** | **Slight:** 0.85<br>**Low:** 1.00<br>**Moderate:** 0.85 | **Slight:** 0.83<br>**Low:** 1.00<br>**Moderate:** 0.92 | **Slight:** 0.88<br>**Low:** 1.00<br>**Moderate:** 0.71 | **Slight:** 0.78<br>**Low:** 1.00<br>**Moderate:** 0.83 | **Slight:** 0.82<br>**Low:** 1.00<br>**Moderate:** 0.77 | **Slight:** 0.15<br>**Low:** 0.00<br>**Moderate:** 0.15 |
| **SupportVector Classifier Model** | **Slight:** 0.95<br>**Low:** 1.00<br>**Moderate:** 0.95 | **Slight:** 0.92<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 0.86 | **Slight:** 0.89<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 0.94<br>**Low:** 1.00<br>**Moderate:** 0.92 | **Slight:** 0.05<br>**Low:** 0.00<br>**Moderate:** 0.05 |
| **Gaussian NB Classification Model** | **Slight:** 0.70<br>**Low:** 0.90<br>**Moderate:** 0.60 | **Slight:** 0.92<br>**Low:** 0.93<br>**Moderate:** 0.54 | **Slight:** 0.38<br>**Low:** 0.80<br>**Moderate:** 0.71 | **Slight:** 0.75<br>**Low:** 0.80<br>**Moderate:** 0.45 | **Slight:** 0.50<br>**Low:** 0.80<br>**Moderate:** 0.56 | **Slight:** 0.30<br>**Low:** 0.10<br>**Moderate:** 0.40 |
| **DecisionTree Classifier Model** | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 0.00<br>**Low:** 0.00<br>**Moderate:** 0.00 |
| **RandomForest Classifier Model** | **Slight:** 0.95<br>**Low:** 1.00<br>**Moderate:** 0.95 | **Slight:** 0.92<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 1.00<br>**Low:** 1.00<br>**Moderate:** 0.86 | **Slight:** 0.89<br>**Low:** 1.00<br>**Moderate:** 1.00 | **Slight:** 0.94<br>**Low:** 1.00<br>**Moderate:** 0.92 | **Slight:** 0.05<br>**Low:** 0.00<br>**Moderate:** 0.05 |

Table 2. Tabular Representation of Each of the Respective Model's Performance Metrics

### 5.2.6. Error Rate

=> It refers to a measure the provides us with an overall fraction of incorrect predictions made by the particular model.

=> Formula:- $\frac{FP + FN}{TP + TN + FP + FN}$

## VI. Conclusion and recommendations

=> Herein, we can observe that Decision Tree Classifier performs the best out of all the other respective models used in all the aspects of the performance metrics. On the other hand, we can see that, the Random Forest Classifier and the Support Vector Classifier models also tend to offer a strengthening balance in the aspects of accuracy, precision and recall, making them suitable for deployment, but only for the "Low chance" category. Furthermore, the Logistic

Regression model also depicts a promising outcome, requiring a little more optimization with respect to the "Moderate chance" category. Thus, will recommend to prioritize the Random Forest Classifier or the Support Vector Classifier, while considering additional insights or validation in case of prediction & not totally be dependent on the Decision Tree Model to confirm its generalization ability in future predictions of cardiac arrests.

## VIII. Future Scope

=> Since, there is no chance of overfitting or imbalanced prediction in the case of Decision Tree Classifier, but, still efforts should be made to make use of ensemble techniques, thus, integrating strengths of multiple models for boosting overall accuracy of the predictions being made. Furthermore, I added the features as per what I knew, but, one can incorporate more new and specific features to improve the reliability of the performance metrics as well as the predictions being carried out. Moreover, one if experienced enough can make use of SHAP (Shapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) in optimizing the respective model's interpretability, thereby, allowing deeper and more reliable insights. Last but not least, one can also make use of continuous learning models to ensure that the models remain relevant and accurate as the years passes by, there is only the need to add the new features and the models can provide the same reliable accuracy or a more accurate one.



## References

**[1].** Targher, G., Day, C. P., & Bonora, E. (2010). Risk of Cardiovascular Disease in Patients with Nonalcoholic Fatty Liver Disease. *New England Journal of Medicine*, *363*(14), 1341–1350. https://doi.org/10.1056/nejmra0912063

**[2].** Lonardo, A., Nascimbeni, F., Mantovani, A., & Targher, G. (2018). Hypertension, diabetes, atherosclerosis and NASH: Cause or consequence? *NAFLD: Emerging Perspectives*, *68*(2), 335–352. https://doi.org/10.1016/j.jhep.2017.09.021

**[3].** METROPOLIS - The Pathology Specialist. (n.d.). *SGPT Test – Normal Range, Uses, Results & More – Metropolis Healthcare*. Metropolis India Lab. https://www.metropolisindia.com/blog/preventive-healthcare/understanding-the-sgpt-test-uses-results-and-normal-range

**[4].** scikit-learn. (2014). *sklearn.linear_model.LogisticRegression — scikit-learn 0.21.2 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

**[5].** Sawant, N., & Khadapkar, D. R. (2022). Comparison of the performance of GaussianNB Algorithm, the K Neighbors Classifier Algorithm, the Logistic Regression Algorithm, the Linear Discriminant Analysis Algorithm, and the Decision Tree Classifier Algorithm on same dataset. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 1654–1665. https://doi.org/10.22214/ijraset.2022.48311

**[6].** Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. In C. Liu, L. Wang, & A. Yang (Eds.), *Information Computing and Applications* (pp. 179–186). Springer Berlin Heidelberg.

**[7].** Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688