IME-672A
Data Mining & Knowledge Discovery

Corporate Rating
Group Number: 4

GROUP MEMBERS:

Pruthviraj Indrajit Desai	***180563***
Pushpanjali Kumari	***180569***
Ravi Kumar	***180594***
Shivam Kumar Vishwakarma	***180719***
Srajal Agrawal	***180785***
Umang Pandey	***180833***

Acknowledgements-

We wish to express our sincere gratitude to our instructor **Dr. Faiz Hamid** for his valuable advice and guidance in completing this project. The way he presented each and every topic in the class made the topics very interesting and understandable, which helped a lot in making our project possible.

Problem Description -

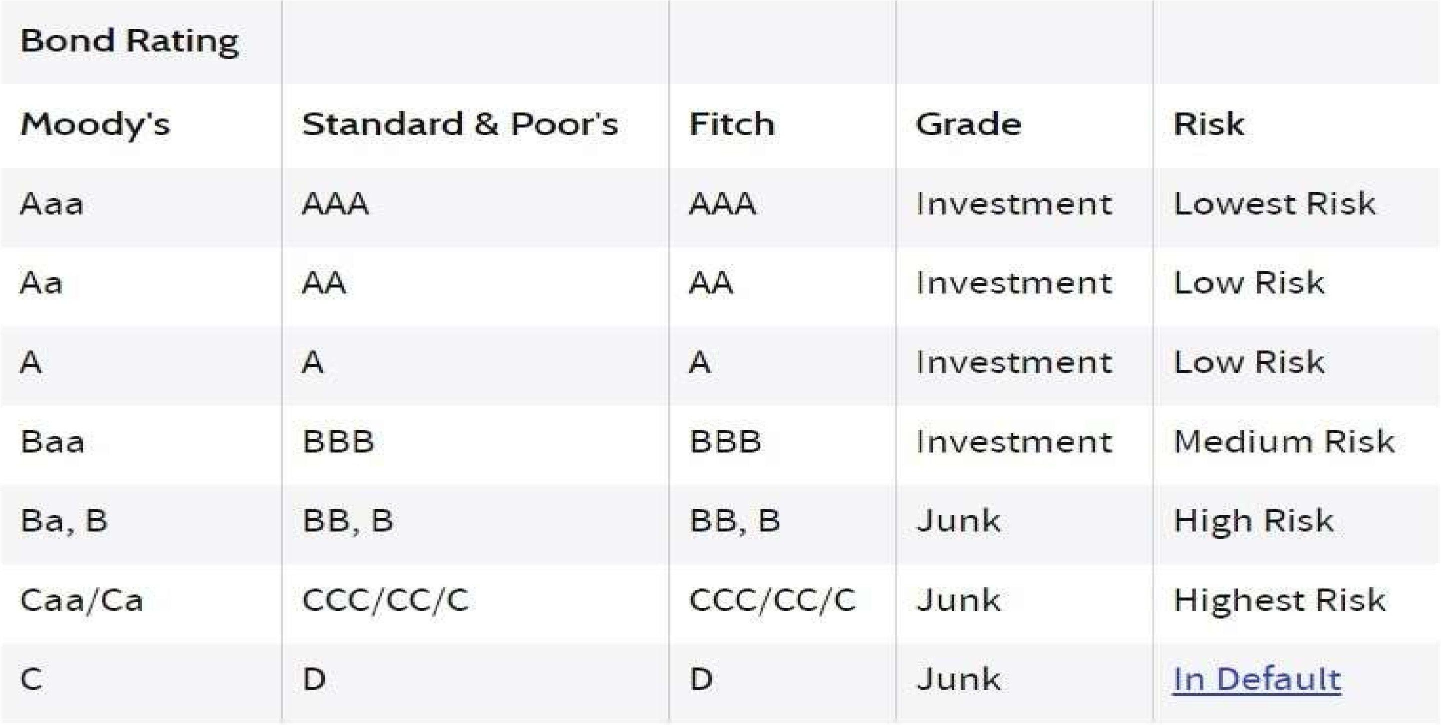
What is a Corporate Credit Rating?

A corporate credit rating is an opinion of an independent agency regarding the likelihood that a corporation will fully meet its financial obligations(meet the terms of a contract) as they come due. A company's corporate credit rating indicates its relative ability to pay its creditors. It is important to keep in mind that corporate credit ratings are an opinion, not a fact.

Key Takeaways

- Corporate credit ratings are the assessment of a company's ability to pay its debts according to an independent credit rating agency.
- The three biggest credit rating agencies are: Standard and Poor's (S&P), Moody's, and Fitch.
- Corporate credit rating trends, over time, may allow an investor to compare the credit-worthiness of competing corporations.

For example, Standard & Poor's uses "AAA" for the highest credit quality with the lowest credit risk, "AA" for the next best, followed by "A," then "BBB" for satisfactory credit.



Bond Rating				
Moody's	Standard & Poor's	Fitch	Grade	Risk
Aaa	AAA	AAA	Investment	Lowest Risk
Aa	AA	AA	Investment	Low Risk
A	A	A	Investment	Low Risk
Baa	BBB	BBB	Investment	Medium Risk
Ba, B	BB, B	BB, B	Junk	High Risk
Caa/Ca	CCC/CC/C	CCC/CC/C	Junk	Highest Risk
C	D	D	Junk	In Default

Description of the Data -

The ratings data set is an anonymized data set with corporate ratings where the ratings have been numerically encoded (1 = AAA, and so on). It has the following attributes:

- **Spid: ID number**
  type - Nominal

- **Rating:**
  type - Ordinal
  range - 1 to 10

- **COMMEQTA: (Common equity to total assets)**
  Common equity is the amount that all common shareholders have invested in a company
  type - Ratio Numeric , continuous

- **LLPLOANS: (Loan loss provision to total loans) -**

A loan loss provision is an income statement expense set aside to allow for uncollected loans and loan payments
type - Ratio Numeric , continuous

- **COSTTOINCOME: (Operating costs to operating income)-**
Operating costs are the ongoing expenses incurred from the normal day-to-day of running a business.Operating income reports the amount of profit realized from a business's ongoing operations.
type - Ratio Numeric , continuous

- **ROE: (Return on equity)**
Return on equity is a measure for financial performance calculated by dividing net income by shareholder’s equity
type - Ratio Numeric , continuous

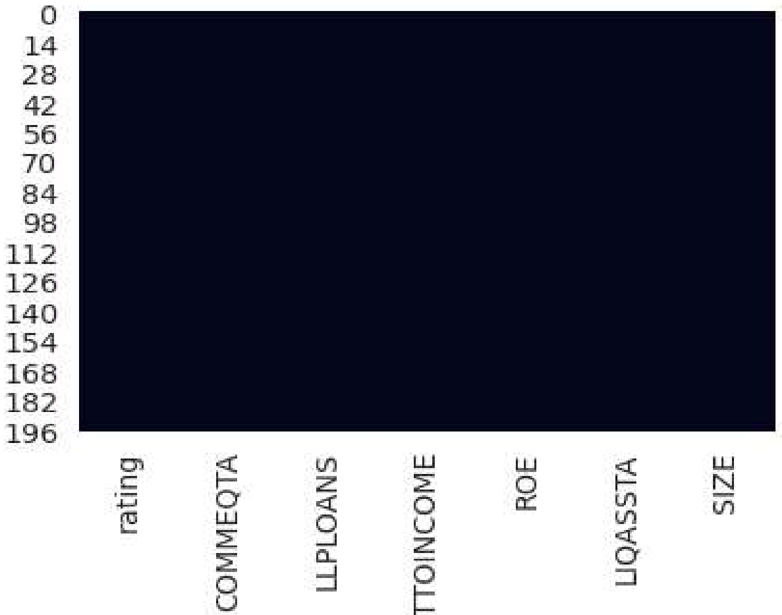
- **LIQASSTA: (Liquid assets to total assets)**
A liquid asset is something you own that can quickly and simply be converted into cash while retaining its market value.
type - Ratio Numeric , continuous

- **SIZE: (Natural logarithm of total assets)**
type - Ratio Numeric , continuous

FEATURES OF DATASET :

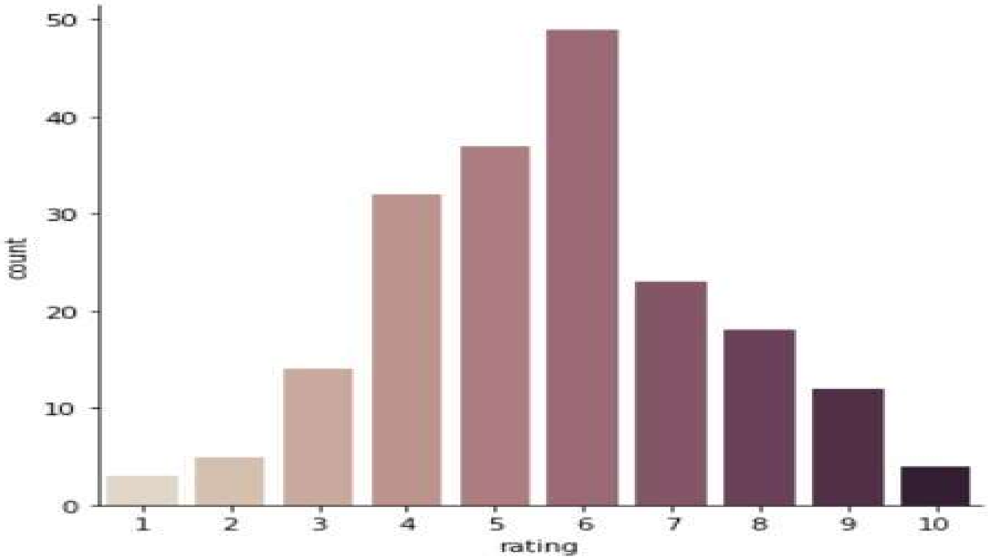
Data Preprocessing

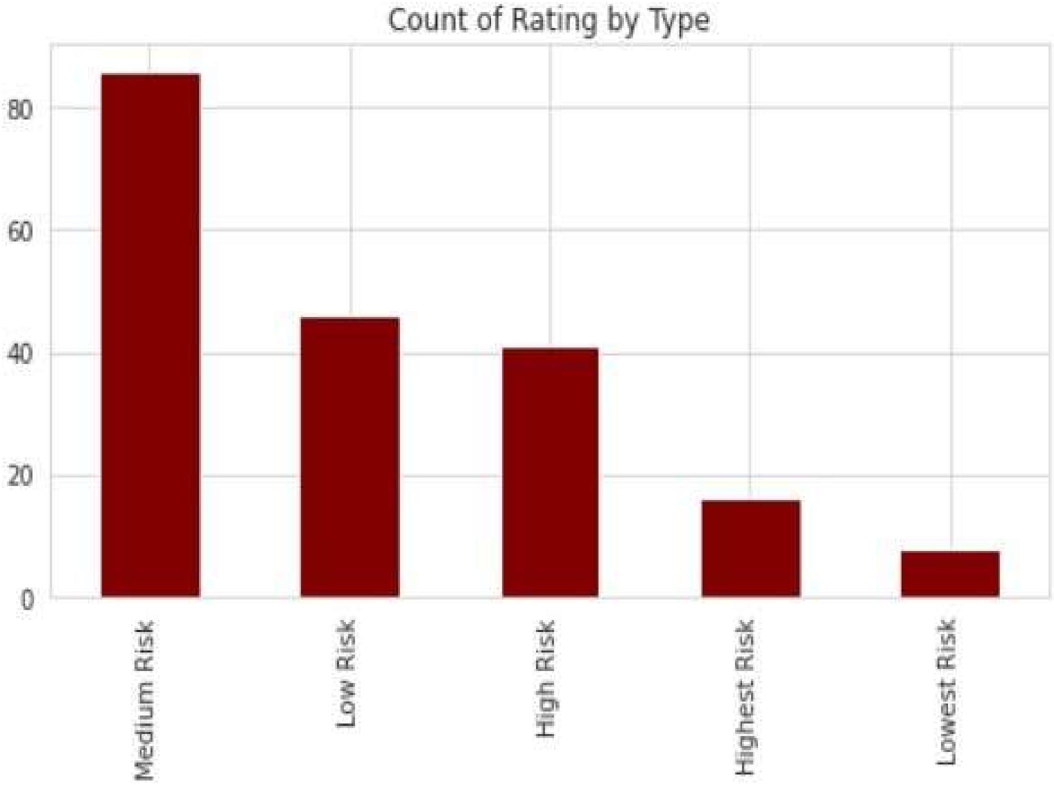
- There is neither any missing value nor any noisy data in our dataset.
- We have scaled ratings between 1-10 with categories as Lowest to highest risk.



Data Visualisation -

This plot shows medium risk rating count is more than others

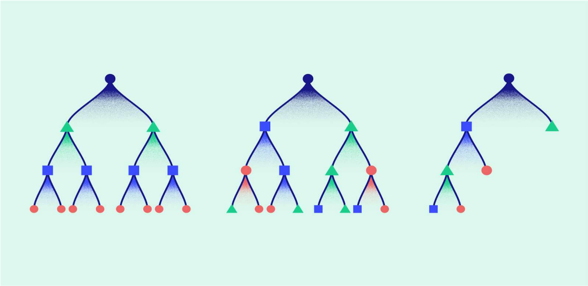




Model Building -

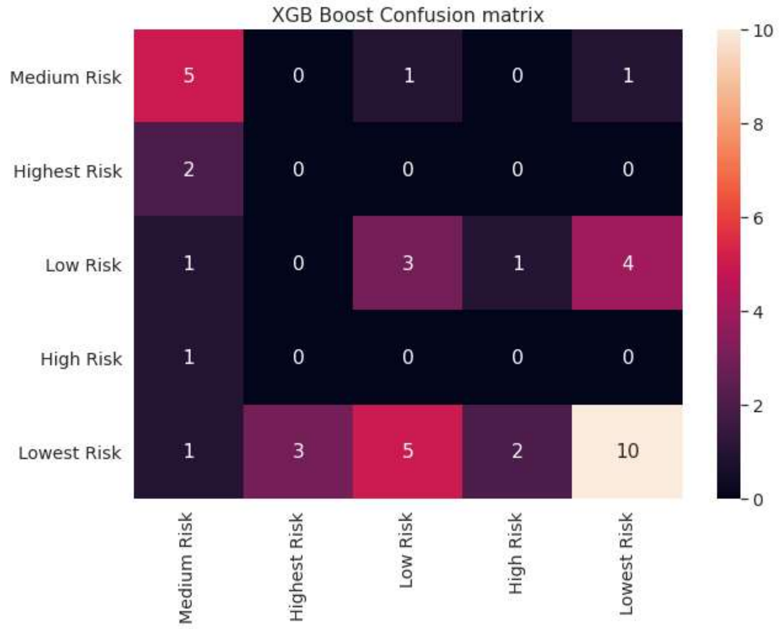
After visualizing and preprocessing the data, we split the data into training and testing data using stratified sampling, splitting it in a 4:1 ratio. We used the training dataset for training 11 models: XGBoost, Gradient Boosted Tree Regression model, Random forest, Support vector machine, Multilayer perceptron, Gaussian Naive bayes, latent dirichlet allocation, Qualitative Data Analysis, K-Nearest Neighbor, Linear regression, Decision tree classifier, Decision tree regressor.

1. XGBoost -



XGB is a popular supervised learning algorithm. It is used in supervised learning in ML. It is an additive and sequential   model which converts weak learners into stronger ones by adding weights to them.

XGB Accuracy:	0.45

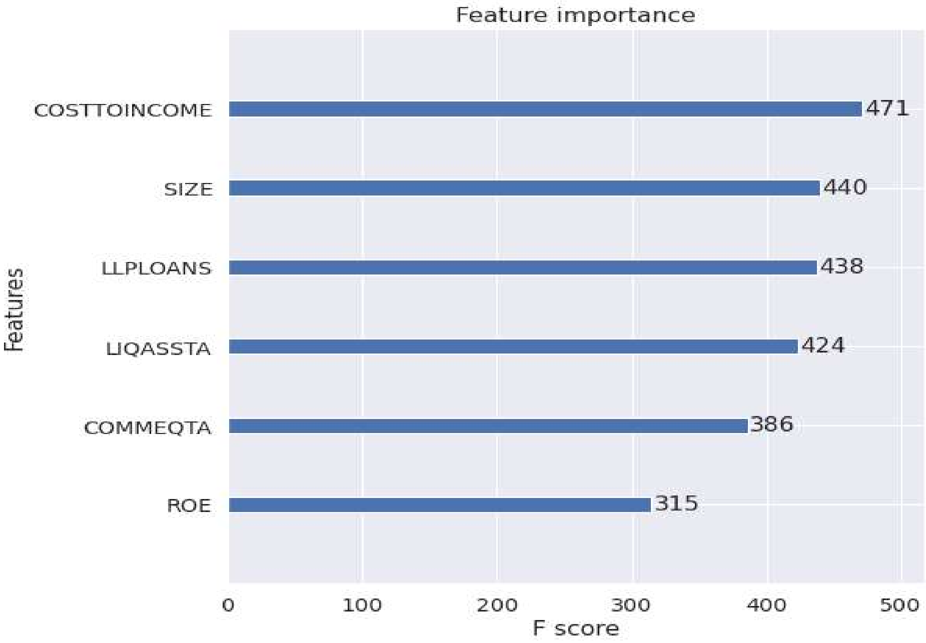


XGB Boost Confusion matrix

	Medium Risk	Highest Risk	Low Risk	High Risk	Lowest Risk
Medium Risk	5	0	1	0	1
Highest Risk	2	0	0	0	0
Low Risk	1	0	3	1	4
High Risk	1	0	0	0	0
Lowest Risk	1	3	5	2	10

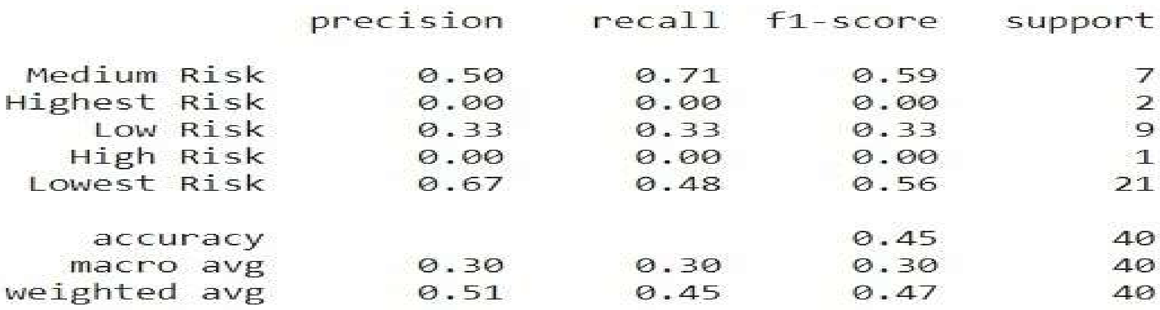
	TP	FP	TN	FN

Medium risk	**5**	**2**	**28**	**5**
Highest risk	**0**	**2**	**35**	**3**
Low risk	**3**	**6**	**25**	**6**
High risk	**0**	**1**	**36**	**3**
Lowest risk	**10**	**11**	**14**	**5**



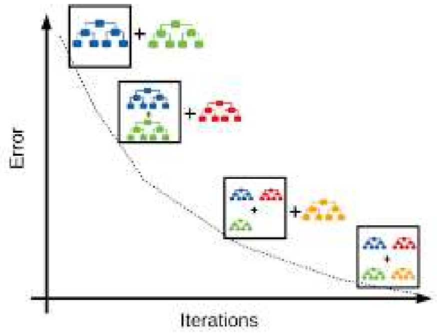
- Operating costs to operating income has the highest value of f-score which shows it’s importance in XBG model. Cost to income is the most important variable towards rating prediction.
- Size of the company is the 2nd most important variable followed by LLPLOANS, COMMEQTA AND ROE.

Classification report

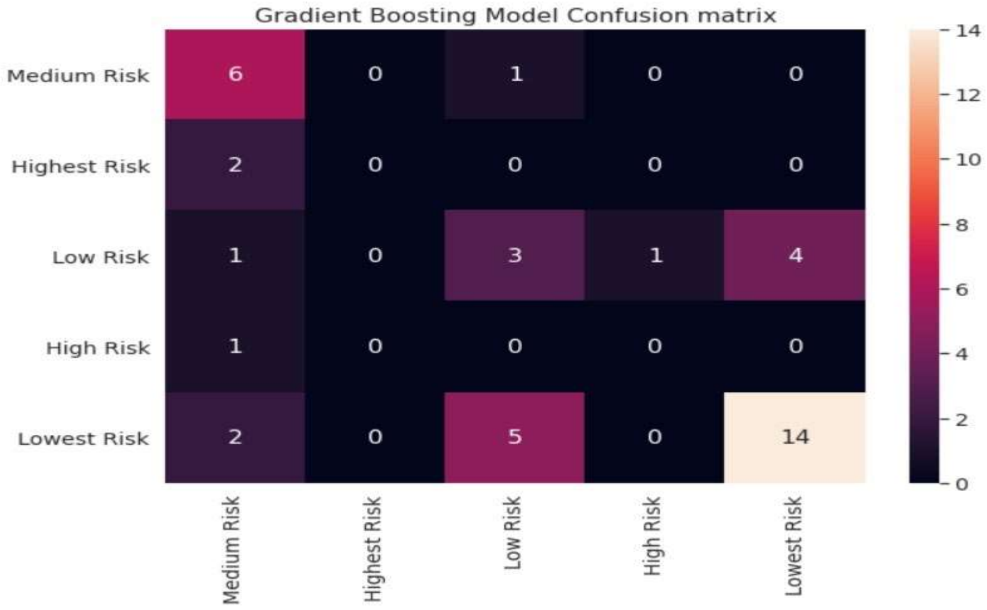


```
                precision    recall  f1-score   support

  Medium Risk        0.50      0.71      0.59         7
 Highest Risk        0.00      0.00      0.00         2
     Low Risk        0.33      0.33      0.33         9
    High Risk        0.00      0.00      0.00         1
  Lowest Risk        0.67      0.48      0.56        21

     accuracy                            0.45        40
    macro avg        0.30      0.30      0.30        40
 weighted avg        0.51      0.45      0.47        40
```

2. GBT -

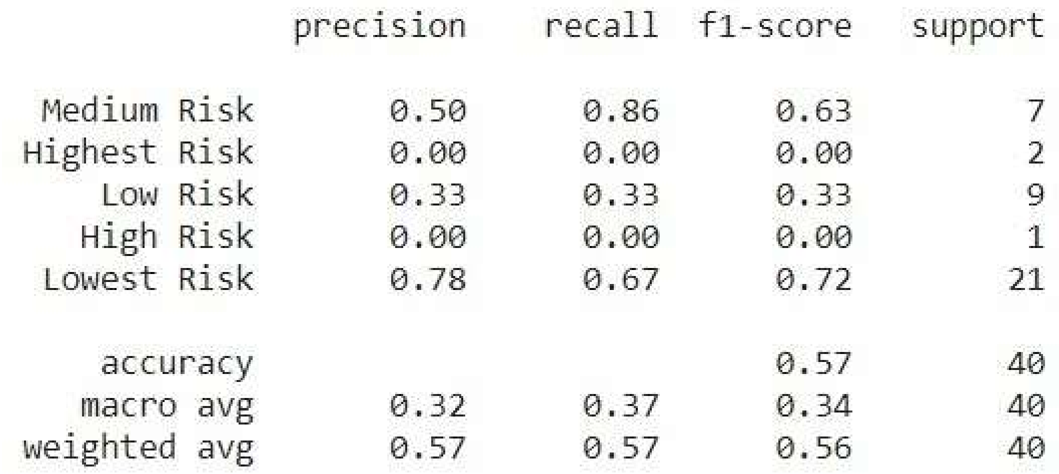


Gradient boosted trees is a learning algorithm for regression. It can be used for classification problems. Since corporate credit rating too is a type of classification problem thus we used GBT to predict the ratings.



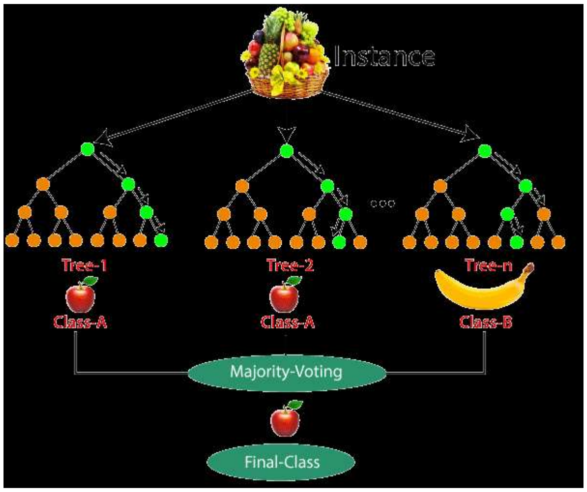
GBT Accuracy:	0.575

Classification report

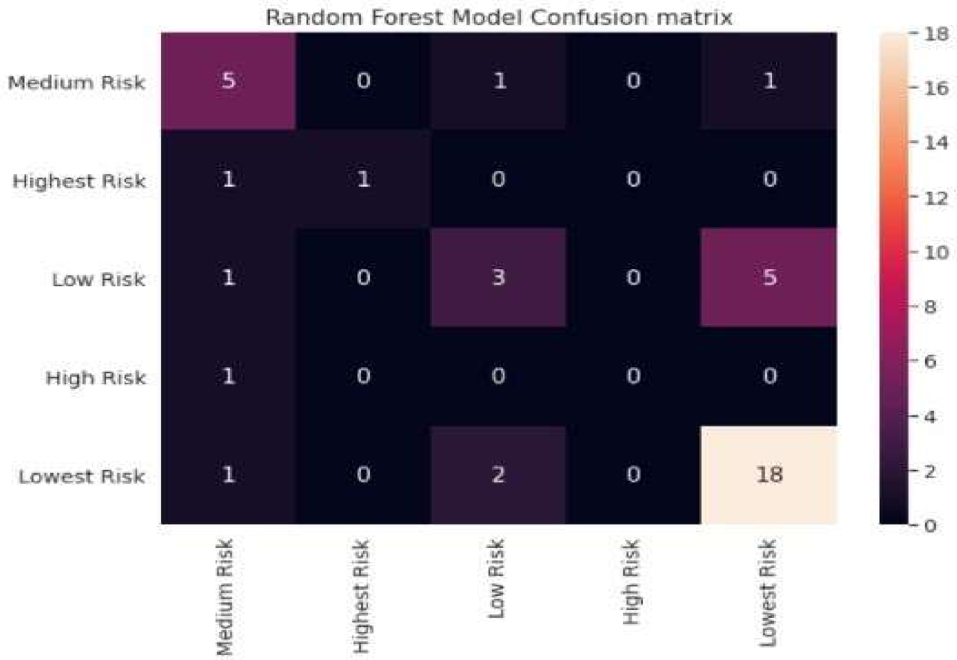


```
                precision    recall  f1-score   support

  Medium Risk       0.50      0.86      0.63         7
 Highest Risk       0.00      0.00      0.00         2
     Low Risk       0.33      0.33      0.33         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.78      0.67      0.72        21

     accuracy                           0.57        40
    macro avg       0.32      0.37      0.34        40
 weighted avg       0.57      0.57      0.56        40
```

3. **Random-Forest (RF) -** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time.RF builds multiple decision trees and merges them together to get a more accurate and stable prediction Since we had to predict the rating like many other algorithms we used this to predict the ratings.

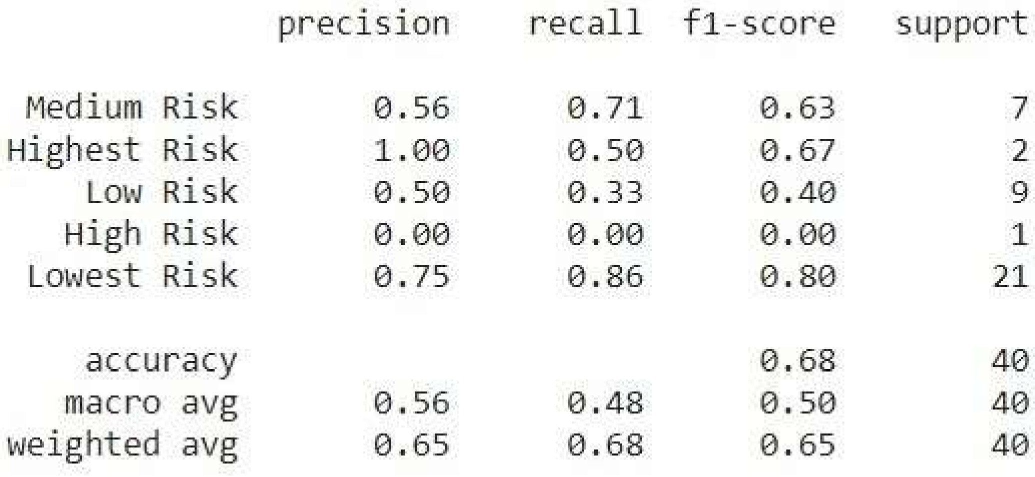


We got maximum accuracy for the RF model.



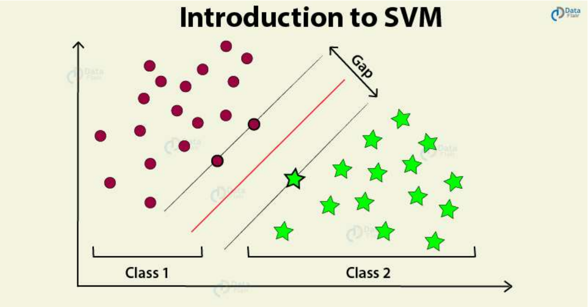
RF Accuracy:	**0.675**

Classification report

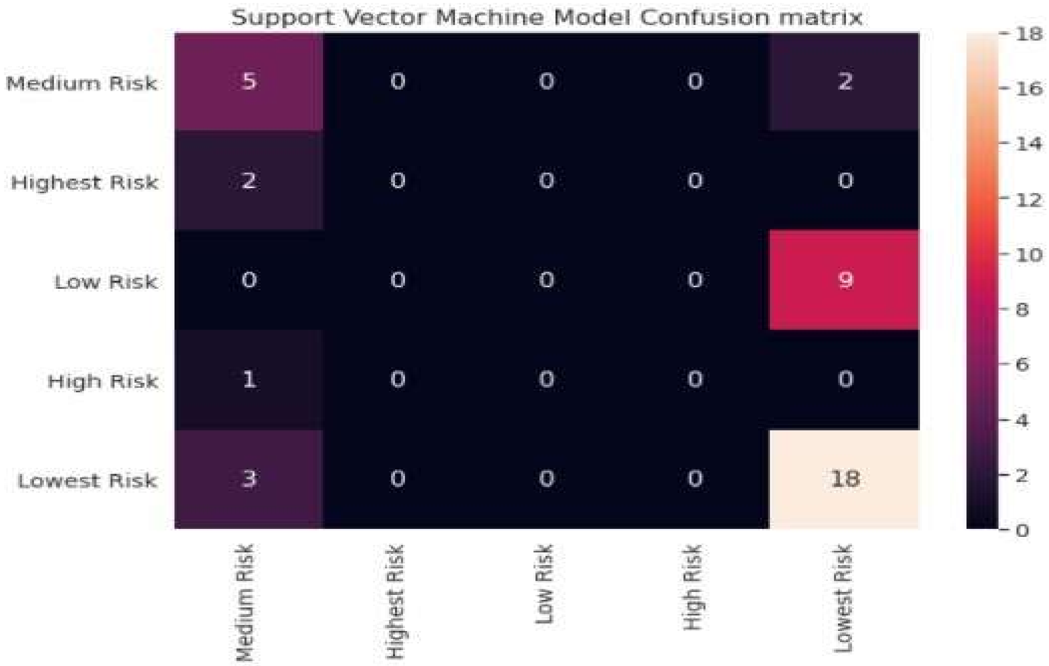


```
               precision    recall  f1-score   support

  Medium Risk       0.56      0.71      0.63         7
 Highest Risk       1.00      0.50      0.67         2
     Low Risk       0.50      0.33      0.40         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.75      0.86      0.80        21

     accuracy                           0.68        40
    macro avg       0.56      0.48      0.50        40
 weighted avg       0.65      0.68      0.65        40
```

4. Support vector machine(SVM) -



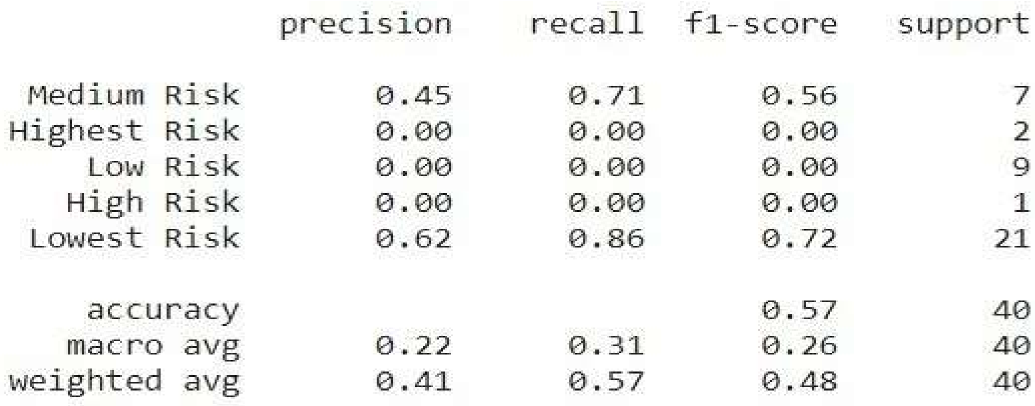
SVM algorithms too can be used for classification problems. SVM works by finding a hyperplane in N-dimensional space that distinctly classifies the data points.


Support Vector Machine Model Confusion matrix

	Medium Risk	Highest Risk	Low Risk	High Risk	Lowest Risk
Medium Risk	5	0	0	0	2
Highest Risk	2	0	0	0	0
Low Risk	0	0	0	0	9
High Risk	1	0	0	0	0
Lowest Risk	3	0	0	0	18

SVM Accuracy:	**0.575**

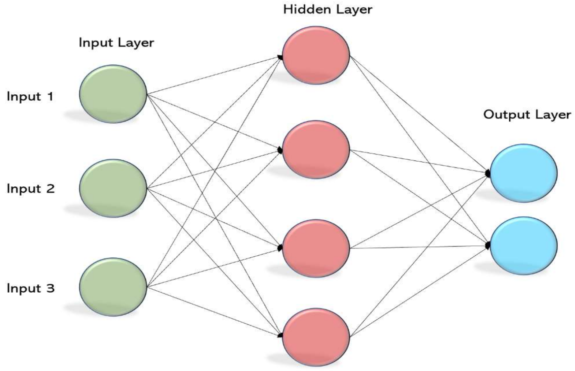
Classification report



```
              precision    recall  f1-score   support

 Medium Risk       0.45      0.71      0.56         7
Highest Risk       0.00      0.00      0.00         2
    Low Risk       0.00      0.00      0.00         9
   High Risk       0.00      0.00      0.00         1
 Lowest Risk       0.62      0.86      0.72        21

    accuracy                           0.57        40
   macro avg       0.22      0.31      0.26        40
weighted avg       0.41      0.57      0.48        40
```

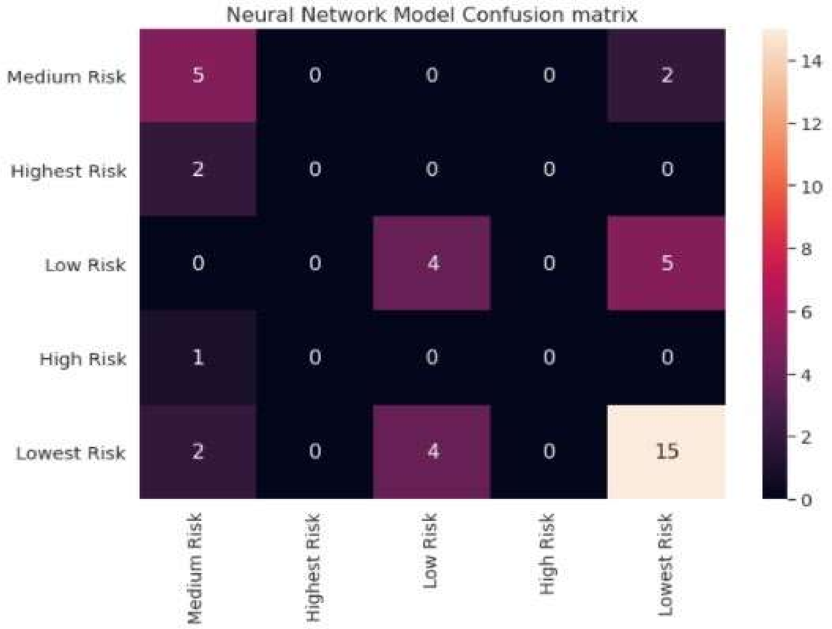
5. Multi-layer Perceptron Classifier (Neural Network) -

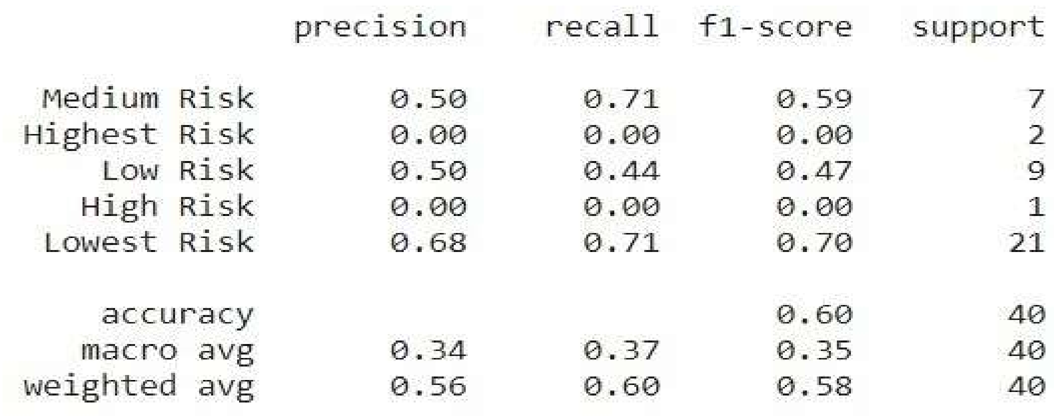
The multilayer perceptron (MLP) is a feedforward artificial neural network model that maps input data sets to a set of appropriate outputs. An MLP consists of multiple layers and each layer is fully connected to the following one. The nodes of the layers are neurons with nonlinear activation functions, except for the nodes of the input layer. Between the input and the output layer there may be one or more nonlinear hidden layers.



Activation function used - f(x) = tanh(x)

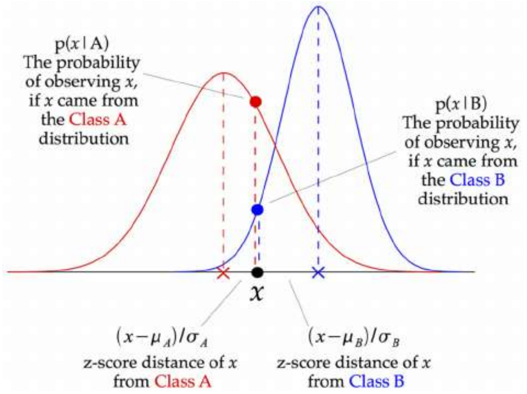
MLP Accuracy:	**0.6**





```
              precision    recall  f1-score   support

 Medium Risk       0.50      0.71      0.59         7
Highest Risk       0.00      0.00      0.00         2
    Low Risk       0.50      0.44      0.47         9
   High Risk       0.00      0.00      0.00         1
 Lowest Risk       0.68      0.71      0.70        21

    accuracy                           0.60        40
   macro avg       0.34      0.37      0.35        40
weighted avg       0.56      0.60      0.58        40
```

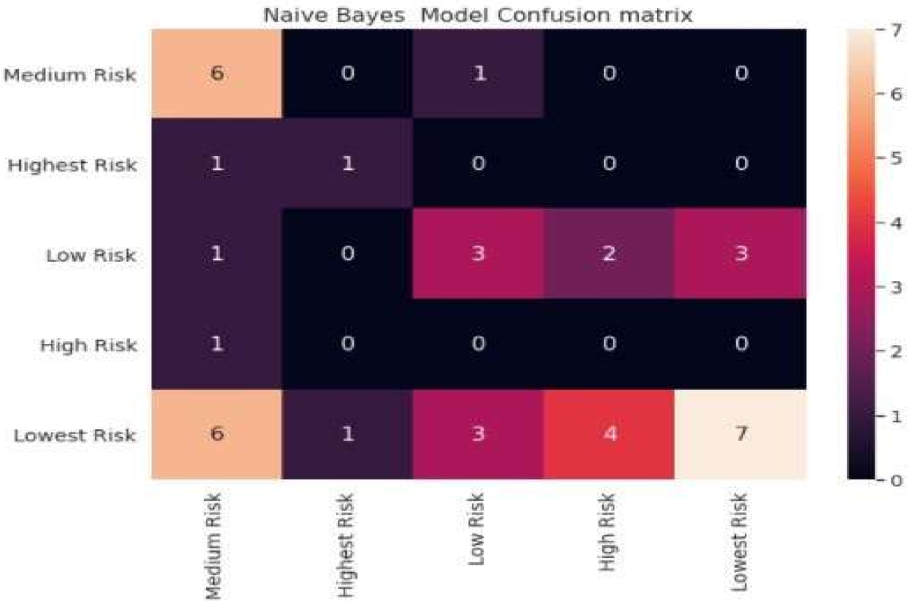
6. Gaussian Naive Bayes Classifier (GNB) -



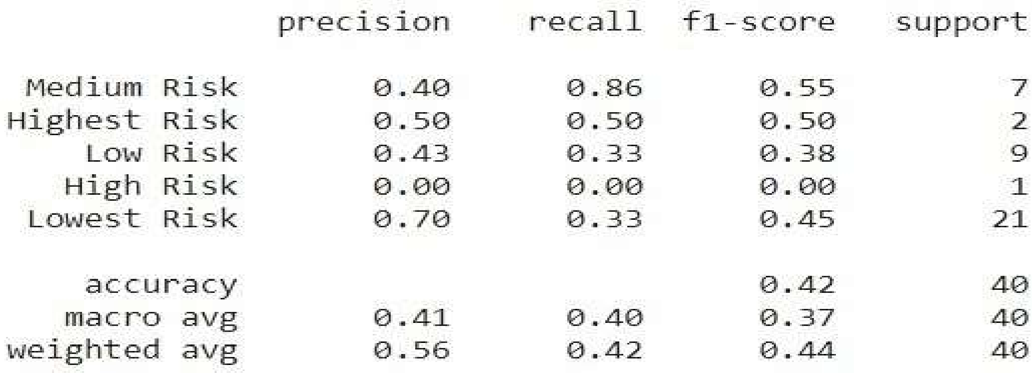
p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. GNB is a basic classification algorithm and we used it just to check how it performs on the given dataset.

GNB Accuracy:	**0.425**



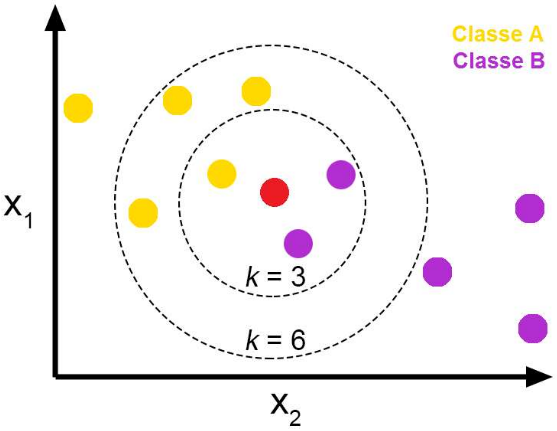
Classification report

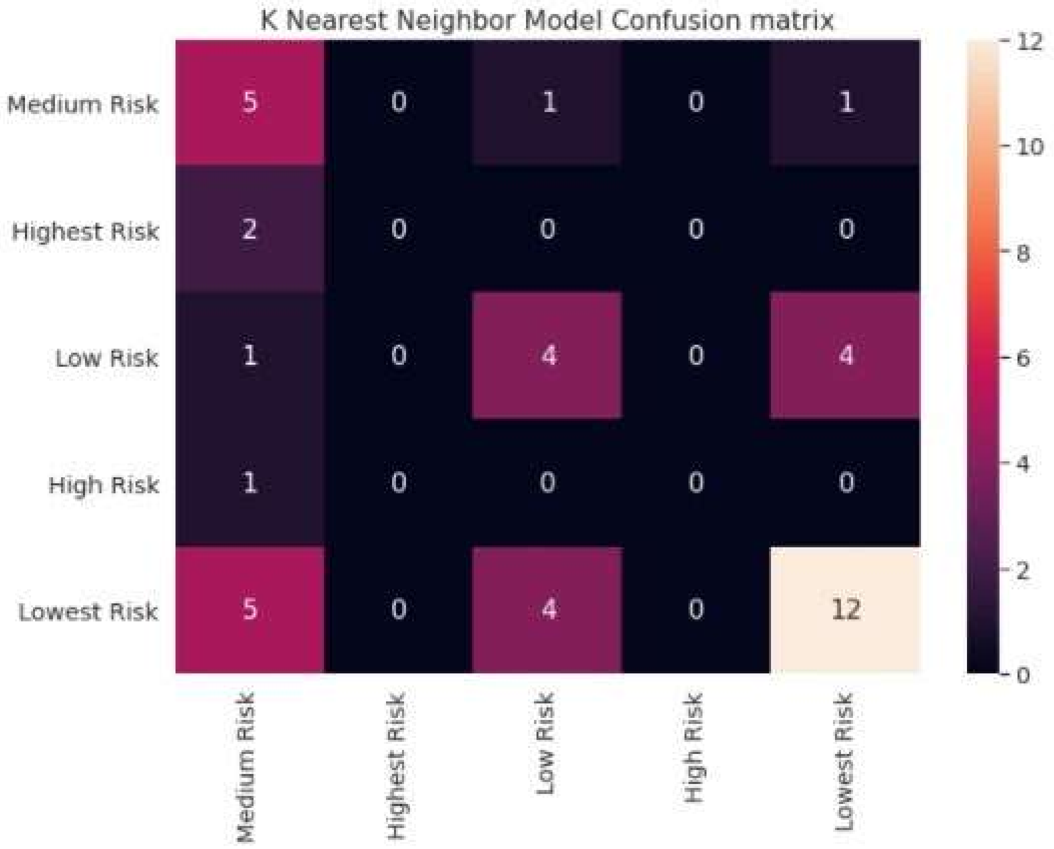


```
              precision    recall  f1-score   support

  Medium Risk       0.40      0.86      0.55         7
 Highest Risk       0.50      0.50      0.50         2
     Low Risk       0.43      0.33      0.38         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.70      0.33      0.45        21

     accuracy                           0.42        40
    macro avg       0.41      0.40      0.37        40
 weighted avg       0.56      0.42      0.44        40
```

7. K-Nearest Neighbours (KNN) -

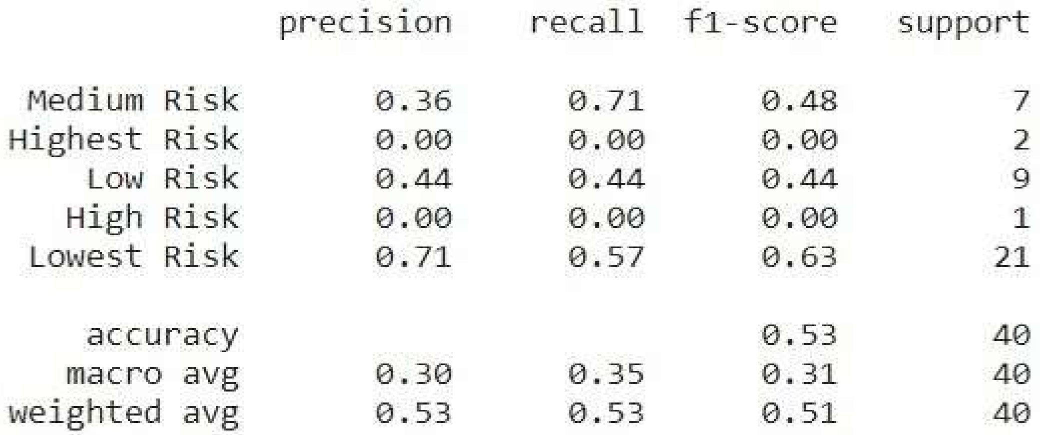
The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slower as the size of that data in use grows. We have specified to use 7 neighbours as default.





KNN Accuracy:	**0.525**

Classification report



```
               precision    recall  f1-score   support

  Medium Risk       0.36      0.71      0.48         7
 Highest Risk       0.00      0.00      0.00         2
     Low Risk       0.44      0.44      0.44         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.71      0.57      0.63        21

     accuracy                           0.53        40
    macro avg       0.30      0.35      0.31        40
 weighted avg       0.53      0.53      0.51        40
```

8. Logistic Regression Model (LR) -

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. We used Multinomial Logistic Regression as it can model scenarios where there are more than two possible discrete outcomes. In the algorithm we used newton-cg solver which uses newton's method to compute the second derivatives.

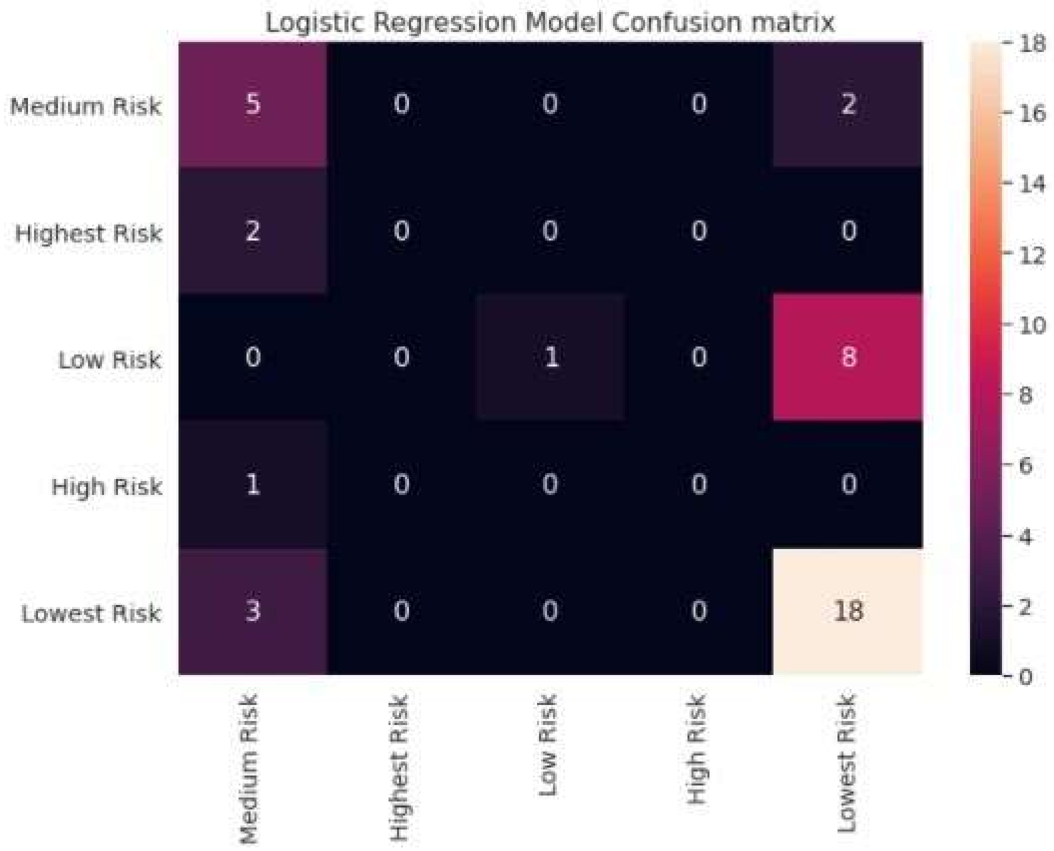
LR Accuracy:	**0.6**

Limitations of Logistic Regression

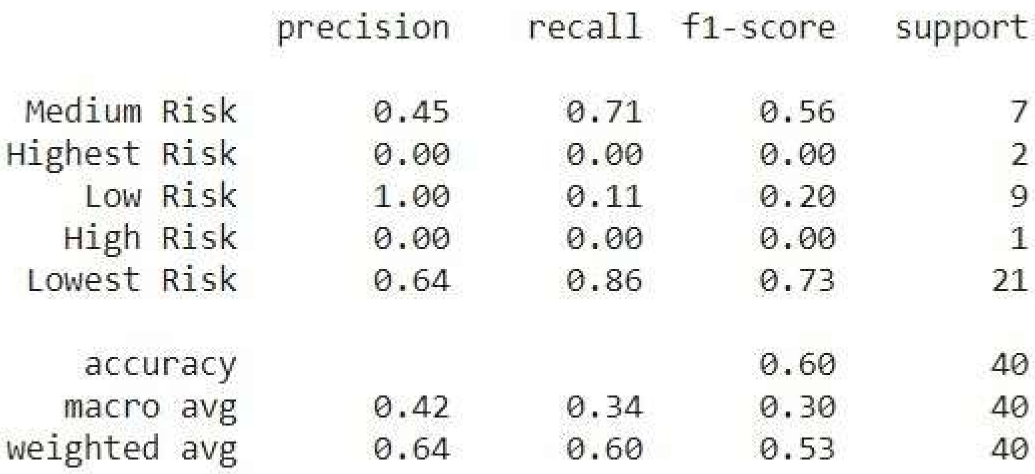
Logistic regression is a simple and powerful linear classification algorithm. It also has limitations that suggest the need for alternate linear classification algorithms.

- Two-Class Problems. Logistic regression is intended for two-class or binary classification problems. It can be extended for multi-class classification, but is rarely used for this purpose.
- Unstable With Well Separated Classes. Logistic regression can become unstable when the classes are well separated.
- Unstable With Few Examples. Logistic regression can become unstable when there are few examples from which to estimate the parameters.

Linear Discriminant Analysis does address each of these points and is the go-to linear method for multi-class classification problems.

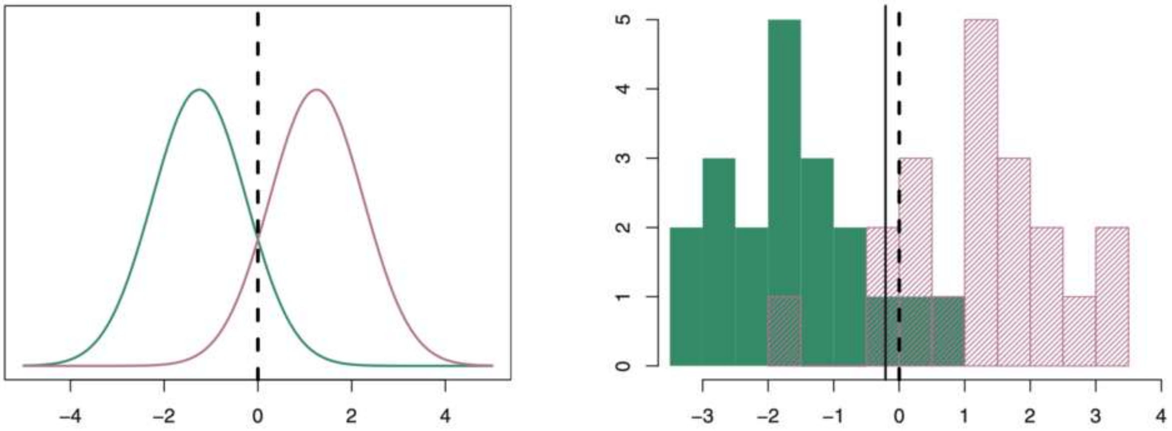


Classification report



```
              precision    recall  f1-score   support

 Medium Risk       0.45      0.71      0.56         7
Highest Risk       0.00      0.00      0.00         2
    Low Risk       1.00      0.11      0.20         9
   High Risk       0.00      0.00      0.00         1
 Lowest Risk       0.64      0.86      0.73        21

    accuracy                           0.60        40
   macro avg       0.42      0.34      0.30        40
weighted avg       0.64      0.60      0.53        40
```

9.Linear Discriminant Analysis :

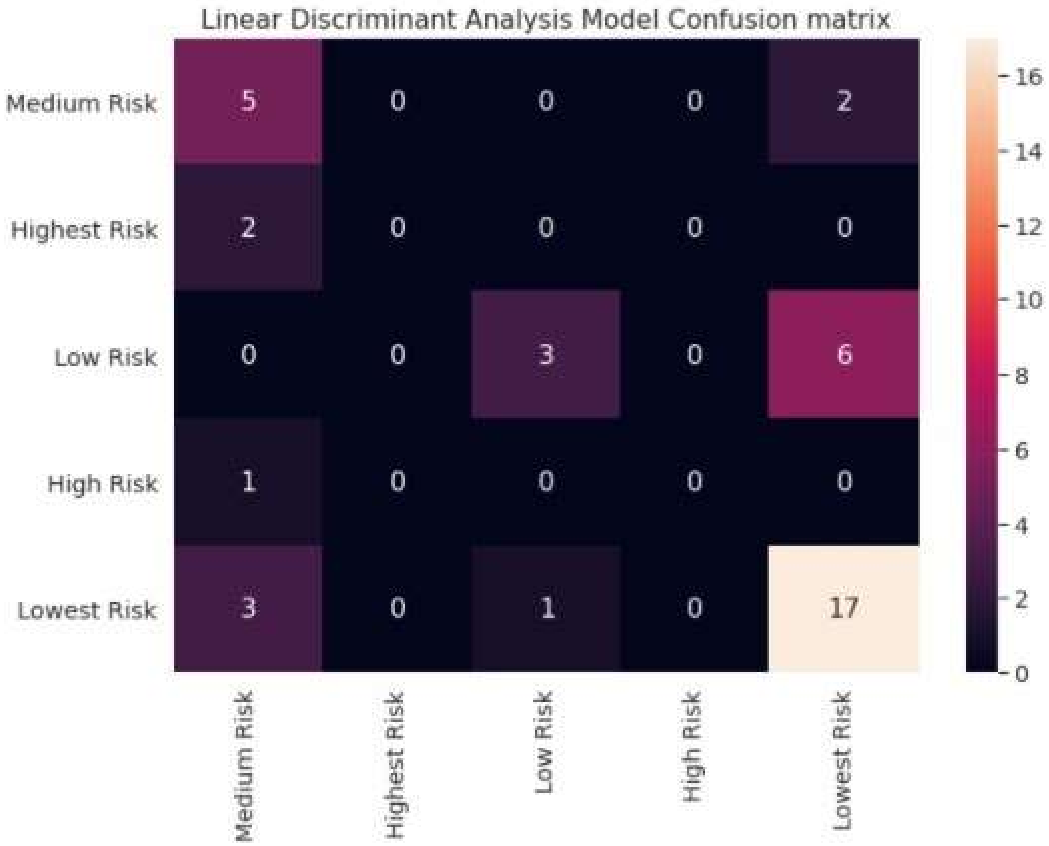


The representation of LDA is straight forward.It consists of statistical properties of your data, calculated for each class. For a single input variable (x) this is the mean and the variance of the variable for each class. For multiple variables, this is the same properties calculated over the multivariate Gaussian, namely the means and the covariance matrix.

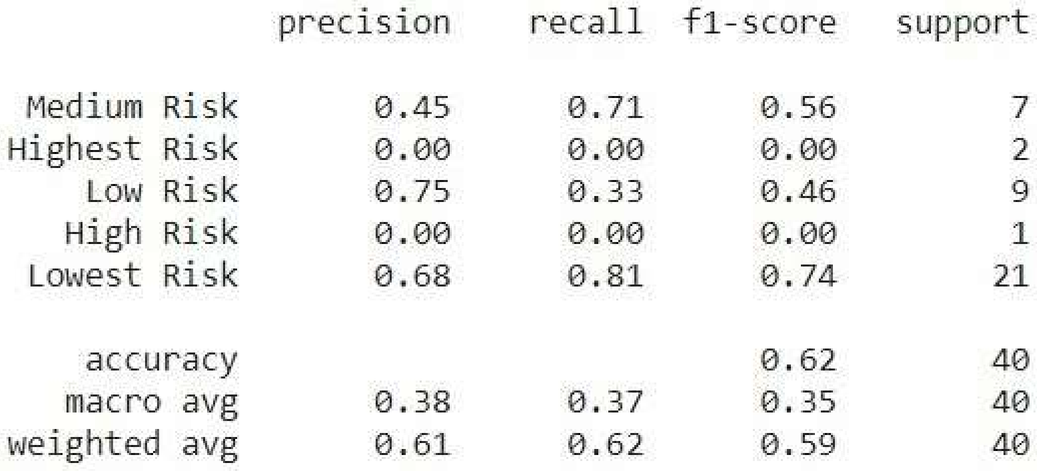
LDA makes some simplifying assumptions about data:

1. That your data is Gaussian, that each variable is is shaped like a bell curve when plotted.
2. That each attribute has the same variance, that values of each variable vary around the mean by the same amount on average.

Since,These Assumptions were really close to what we had Inferred from data analysis we Decided to Implement the model and it ended up being one of the best performers out of the bunch.



Classification report



```
               precision    recall  f1-score   support

  Medium Risk       0.45      0.71      0.56         7
 Highest Risk       0.00      0.00      0.00         2
     Low Risk       0.75      0.33      0.46         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.68      0.81      0.74        21

     accuracy                           0.62        40
    macro avg       0.38      0.37      0.35        40
 weighted avg       0.61      0.62      0.59        40
```

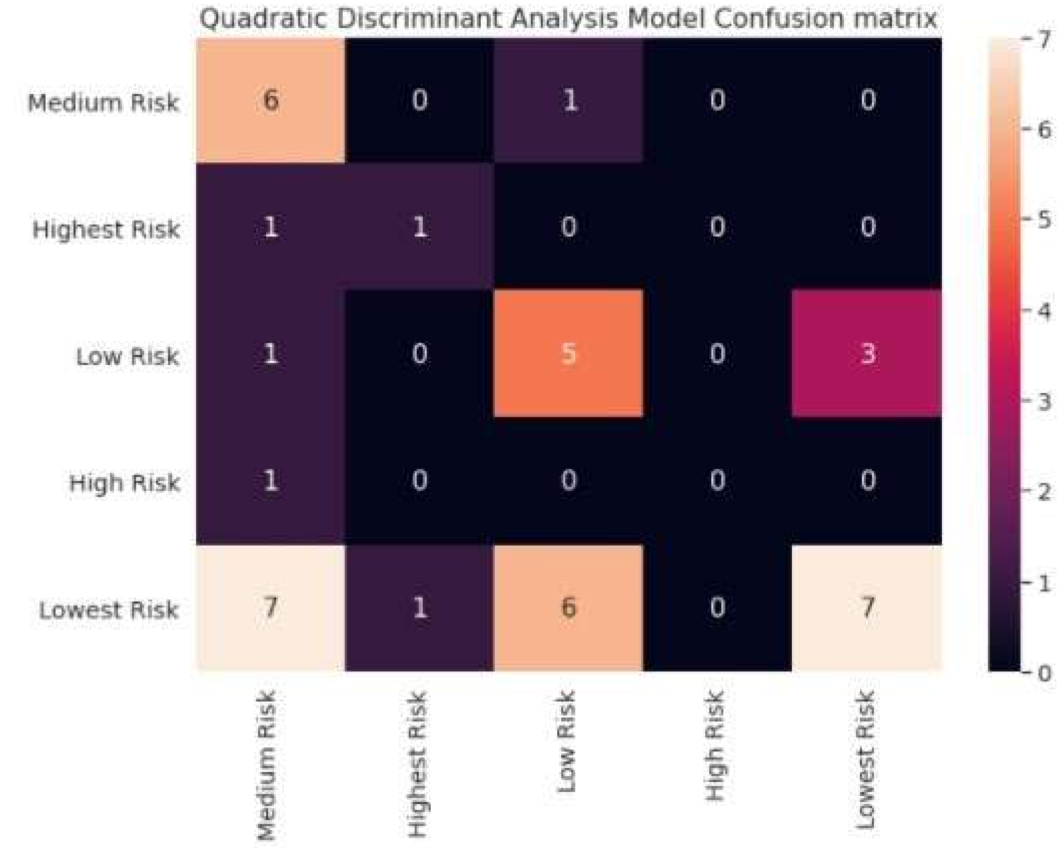
10.Quadratic Dicriminant Analysis :

Quadratic Discriminant Analysis (QDA) is similar to LDA based on the fact that there is an assumption of the observations being drawn form a normal distribution. The difference is that QDA assumes that each class has its own covariance matrix, while LDA does not.

The QDA makes these assumptions about the data :

- Observation of each class is drawn from a normal distribution (same as LDA).
- QDA assumes that each class has its own covariance matrix (Different from LDA).

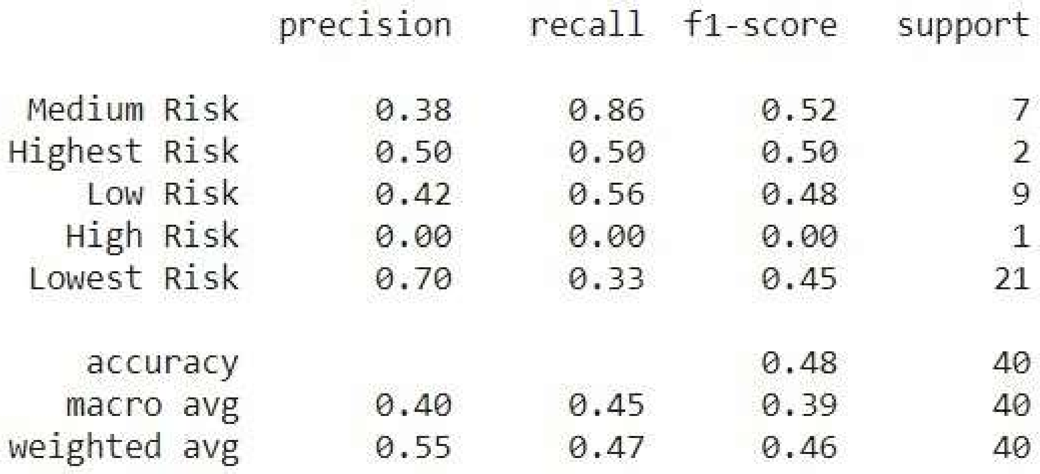
In conclusion LDA is less flexible than QDA because we have to estimate fewer parameters. This can be good when we have only a few observations in our training data which was the case with us.



Quadratic Discriminant Analysis Model Confusion matrix

	Medium Risk	Highest Risk	Low Risk	High Risk	Lowest Risk
Medium Risk	6	0	1	0	0
Highest Risk	1	1	0	0	0
Low Risk	1	0	5	0	3
High Risk	1	0	0	0	0
Lowest Risk	7	1	6	0	7

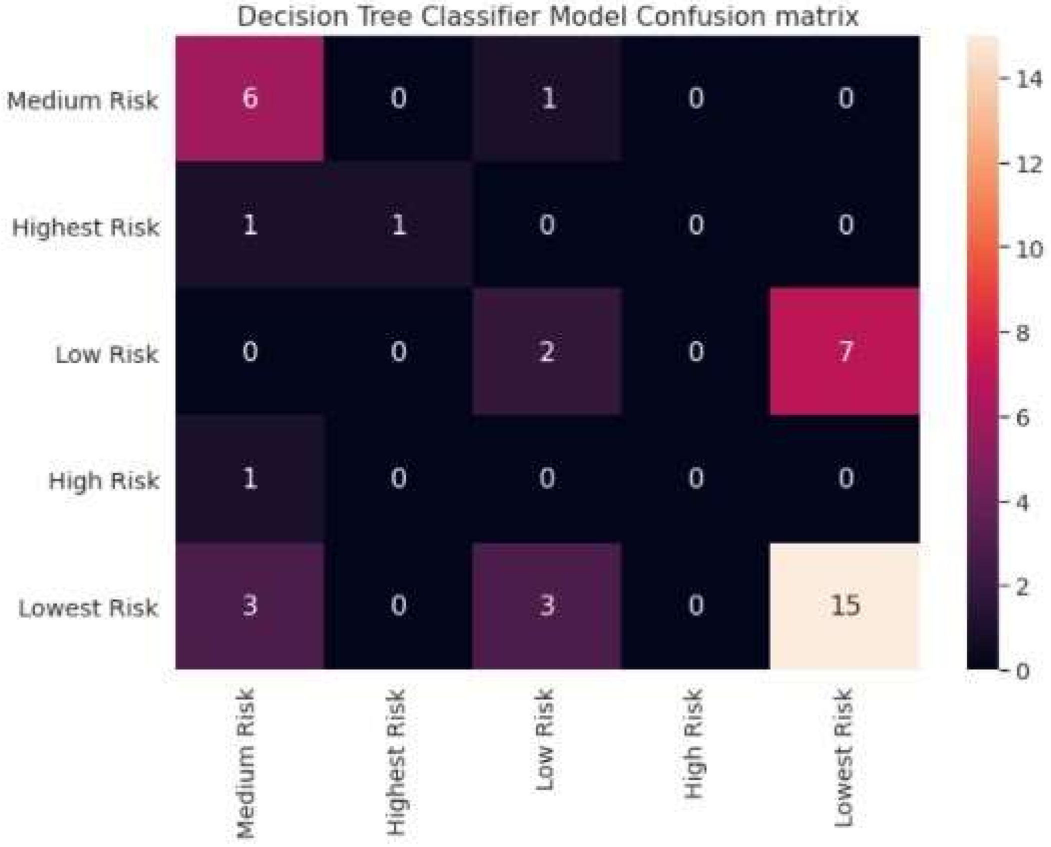
Classification report



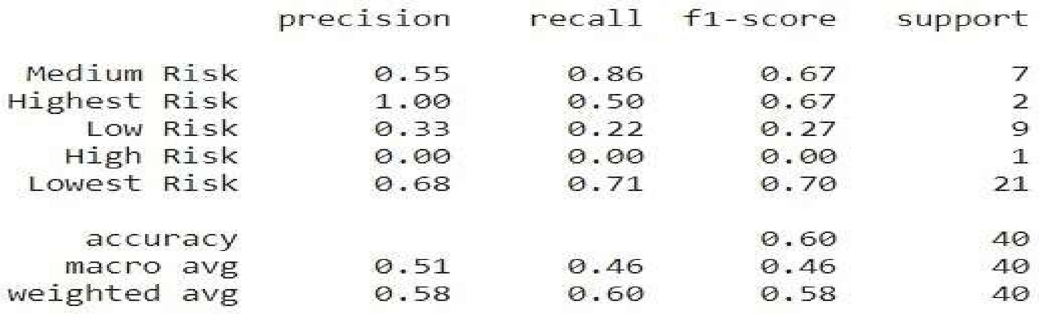
```
               precision    recall  f1-score   support

  Medium Risk       0.38      0.86      0.52         7
 Highest Risk       0.50      0.50      0.50         2
     Low Risk       0.42      0.56      0.48         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.70      0.33      0.45        21

     accuracy                           0.48        40
    macro avg       0.40      0.45      0.39        40
 weighted avg       0.55      0.47      0.46        40
```

11.<u>Decision Tree Classifier</u> :

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

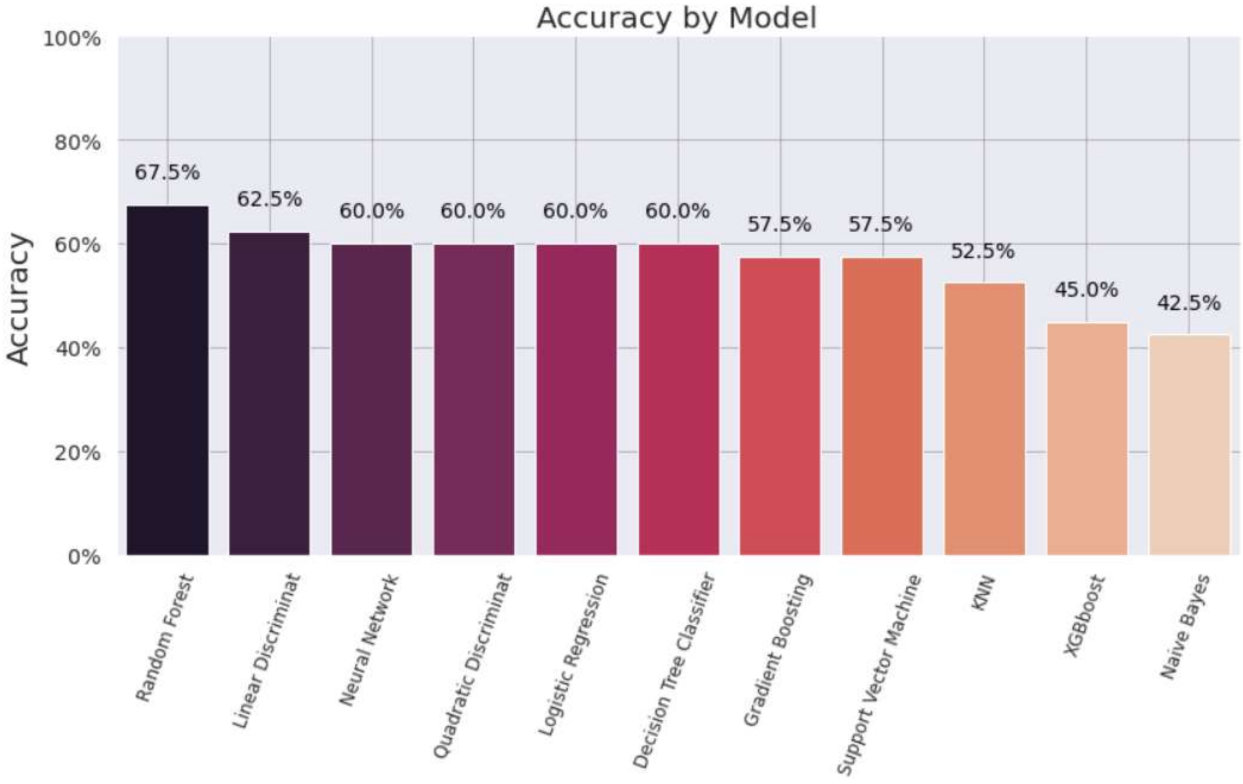


Classification report



```
               precision    recall  f1-score   support

  Medium Risk       0.55      0.86      0.67         7
 Highest Risk       1.00      0.50      0.67         2
     Low Risk       0.33      0.22      0.27         9
    High Risk       0.00      0.00      0.00         1
  Lowest Risk       0.68      0.71      0.70        21

     accuracy                           0.60        40
    macro avg       0.51      0.46      0.46        40
 weighted avg       0.58      0.60      0.58        40
```

Conclusion -



- **XGB Accuracy: 0.45**
- **GBT Accuracy: 0.575**
- **RF Accuracy: 0.675**
- **SVM Accuracy: 0.575**

- **MLP Accuracy: 0.6**
- **GNB Accuracy: 0.425**
- **KNN Accuracy: 0.525**
- **LR Accuracy: 0.6**
- **LDA Accuracy: 0.625**
- **QDA Accuracy: 0.475**
- **Decision Tree classifier: 0.6**

Since , Random-Forest/Decision-Trees Does not inherently make any and/or require any Correlation between attributes as the decision tree is a distribution-free or non-parametric method. Instead it deploys a greedy strategy which instead looks for the best dividing metric possible at that given step , it also iis relatively less data hungry than some of the other models than we used.

Due to the above mentioned reasons the Random Forests ended up performing better than every other model. Followed closely by ,Decision Trees, LDA ,Neural nets and QDA.