A PROJECT REPORT ON

Analysis of Mental Health of Post-Graduation Students in Shivaji University, Kolhapur

SUBMITTED TO

DEPARTMENT OF STATISTICS

SHIVAJI UNIVERSITY

KOLHAPUR.

FOR THE PARTIAL FULFILLMENT OF THE DEGREE

M.Sc. Applied Statistics and Informatics

SUBMITTED BY,

Miss. Vanjale Hrutuja Rajendra

Under the guidance of

Mr. S. M. Patil

DEPARTMENT OF STATISTICS

SHIVAJI UNIVERSITY, KOLHAPUR

2021-2022.

CERTIFICATE

This is to certify that the project entitled **"Analysis of *Mental Health of Post-Graduation Students in Shivaji university, Kolhapur*"** as partial fulfilment for the award of the degree of M.Sc. Applied Statistics and informatics (Part II) of Shivaji University, Kolhapur, is a record of bonafide work carried out by Miss. Vanjale Hrutuja Rajendra under my supervision and guidance.

To the best of my knowledge, the matter presented in the project has not been submitted earlier.

Place: Kolhapur

Date:

Mr. S. M. Patil
Project Guide

Prof. (Dr.) S. B. Mahadik
Head of Department,
Department of Statistics,
Shivaji University, Kolhapur.

PREFACE

It is a great opportunity for me to have the M.Sc. in Applied Statistics and informatics at department of statistics in Shivaji university, Kolhapur. In the accomplishment of this degree, I am submitting a project report on "Analysis of Mental Health of Post-Graduation Students in Shivaji University, Kolhapur." Subject to the limitation of time efforts and resources, every possible attempt has been made to study the problem deeply. The whole project is measured through the primary data, the further analysed and interpreted and the result was obtained.

This project is presented in simple and lucid language. We would fill amply rewarded if the project would prove to be beneficial to anyone who studies it.

ACKNOWLEDGEMENT

I wish to thank the Department of Statistics, Shivaji University, Kolhapur for giving me an opportunity to do the project. This report has been prepared under the guidance of Mr. S. M. Patil. I would like to express my profound gratitude towards him for his guidance and timely support throughout the completion of this project.

Also, I would like to thank Prof.(Dr.) S. B. Mahadik, head of the department of statistics for his support and guidance. I would also like to thank all other faculty members, non-teaching staff and research students for their support, suggestions, and guidance for this project.

I would also like to thank Dr. Pankaj Kupwade (M.B.B.S.,D.C.H.) for spending his valuable time for the questionnaire preparation. I would like to thank all my friends and senior students for their co-operation and help which we received from them throughout the work.

CONTENTS

INTRODUCTION

Mental illness is a health problem that undoubtedly impacts emotions, reasoning, and social interaction of a person. After Covid 19, especially the PG students on the brink of making their careers faced these issues widely. This makes them vulnerable to various psychiatric conditions. Their mental health is affected due to a variety of causes like heredity, attachment patterns with parents, demands and frustrations experienced in schools and colleges, low self-esteem, bodily changes, relationship issues, living in an unsafe or toxic environment, separation or divorce of parents, chronic illness in the family, death of a loved one, moving or changing schools, financial problems, addiction to the internet, etc. These issues have shown that mental illness gives serious consequences across societies and demands new strategies for prevention and intervention. To accomplish these strategies, early detection of mental health problems is an essential procedure. Lack of knowledge about the mental illnesses poses a challenge to the mental health care delivery system. Adolescence is associated with changes in the brain, body and social environment. In some cases, stigma is due to lack of awareness, denying mental health symptoms, poor or inadequate treatment among college students and it contributes to the persistence of mental health problems in this population. UNICEF conducted a 21-nation survey on help-seeking behavior and found out that only 41

per cent of young people in India seek support for mental health problems, compared to an average of 83 per cent for 21 countries. These numbers are concerning.

The country will be taken care of by the young minds. It is essential to focus on their development and wellbeing. No empowerment can happen without psychological empowerment. If they are given appropriate support in different areas of life, including mental health, they will turn into rational, emotionally stable, healthy and productive adults.

By considering all these things I decided to do our project work on this subject. To achieve our claim, it is important to collect the response of the students from Shivaji university, Kolhapur. We collected the data from post-graduation students of this university.

Because post-graduate students are the ones on the verge of making their careers and being a post-graduate student myself, I was curious about my fellow students. Hence, I decided to do the project to see if post graduate students are in need of medical counselling and consultation or not.

OBJECTIVES

- To check that what factors affect students the most that they need a medical counselling and consultation.
- Checking relationship of different explanatory variables with target variable (need of medical consultation and counselling.)
- To check the gender wise need a medical counselling and consultation.
- To build a good predictive model based the collected data.

DATA COLLECTION

For this project, I have collected a primary data through a voluntary response survey conducted via google forms. I used google forms since it is efficient and less tedious compared to traditional survey methods. I shared the link with students via e-mails, messages, etc.

I have collected a data from all three streams (Arts, Commerce and Science) of size 302. The collected data contains variables Disturbed sleeping cycle, Changes in eating habits, Worse personal relations, Loneliness experienced by students, Overthinking, Insecurity in students about the future.

Several extra details were included in the questionnaires to cover all possible topics.

INFORMATION ABOUT VARIABLES

Gender	Gender of the student
Stream	Stream of the student
Disturbed sleeping cycle	If the student has disturbed sleeping cycle
Changes in eating habits	If the student has experienced changes in eating habits
Covid 19 impact	What impact has covid 19 had on them
Loneliness	If the student experienced loneliness
Overthinking	If the student overthinks things or not
Worse personal relations	If the student has had worse personal relations during covid
Worse living condition	If the student has had worse living condition during covid
Insecurity	If the student faces insecurity about the future
Hopelessness	If the student feels hopeless about their life consistently
Need of medical counselling and consultation	If the student is in the need of medical counselling and consultation or not

All the above variables are categorical and have responses in form of Yes and No encoded as 1 and 0 resp. while male is encoded as 0 and female as 1.

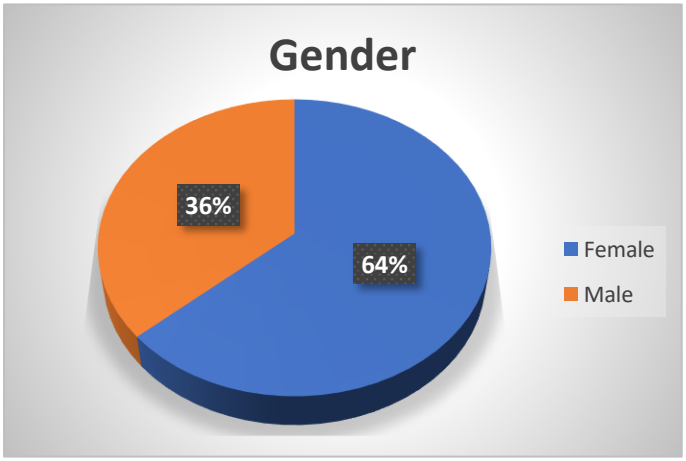
Methodology

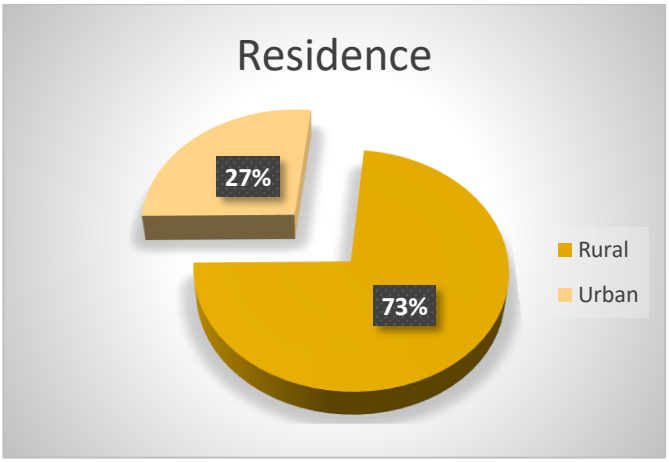
- In this project, the planning phase is conducted followed by the searching and analysis phase. For the planning stage, the research papers and articles are identified.

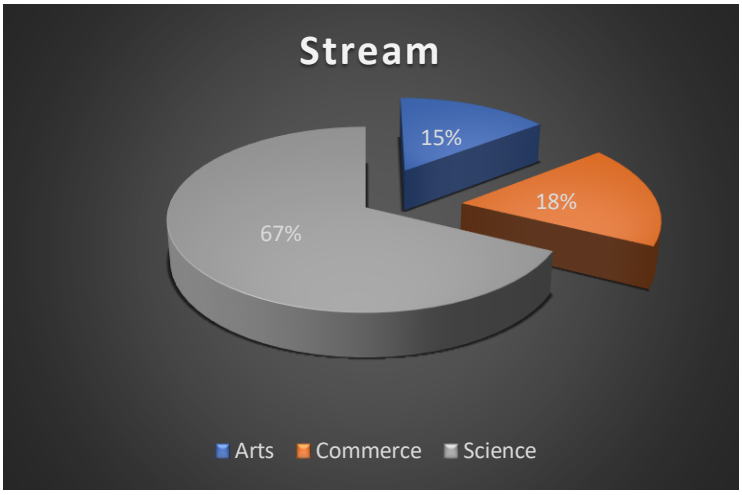
- To conduct the searching and analysis, the topic stated has been explored in the several publishers' website. Besides that, the queries such as Machine Learning Algorithms in Mental Health, and Machine Learning in Predicting Mental Health Problems have been researched for this project.

- The analysis phase is started by finding out and investigating the performance of the machine learning approaches that were used to diagnose or predict mental health problems. For our data, we have compared different classification models with each other with the use of python to see which model gives the better prediction.

Data at Glance







Interpretation: From above pie charts, it is clear that-

- The **female percentage(64%)** is larger than the male in the response survey.
- Maximum percentage (**73%**) of the respondents is from the **rural background.**
- The maximum (**67%**) respondents are from **science stream.**

STATISTICAL ANALYSIS

GRAPHICAL REPRESENTATION

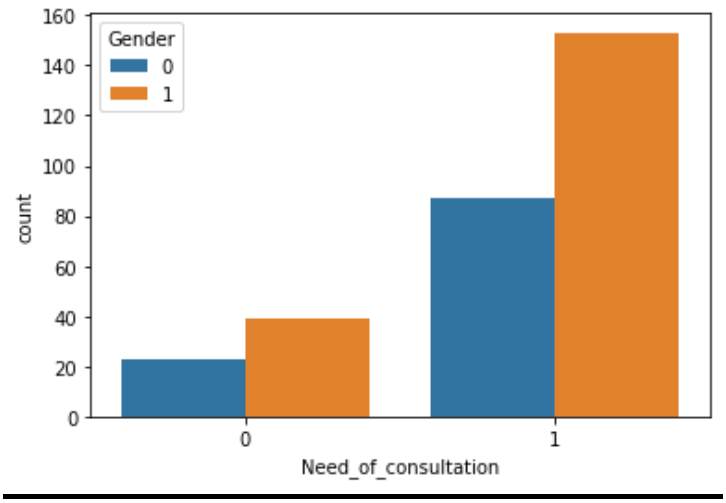


Fig.1.1

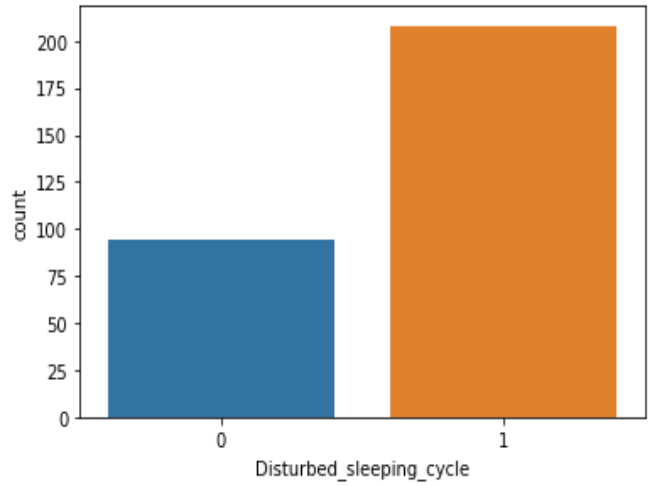


Fig. 1.2

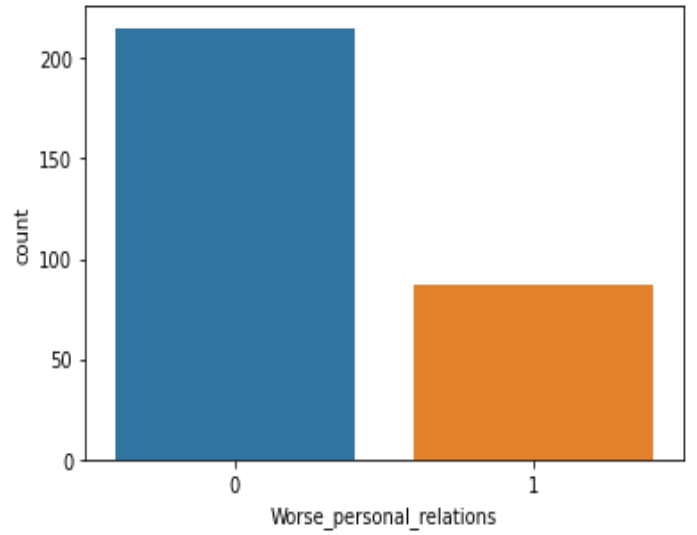


Fig. 1.3

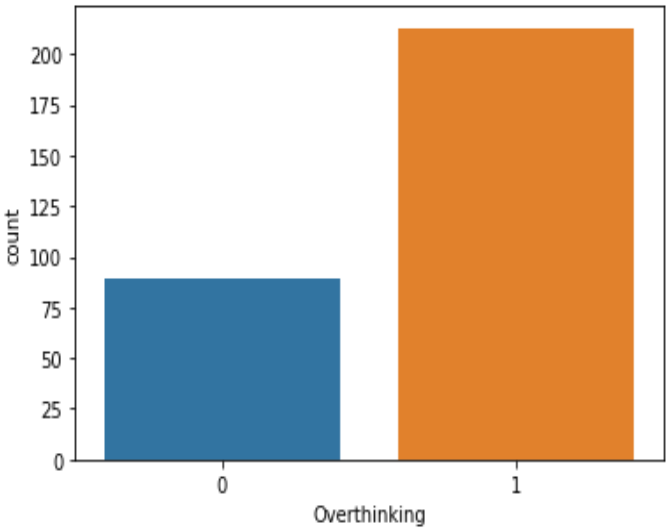


Fig. 1.4

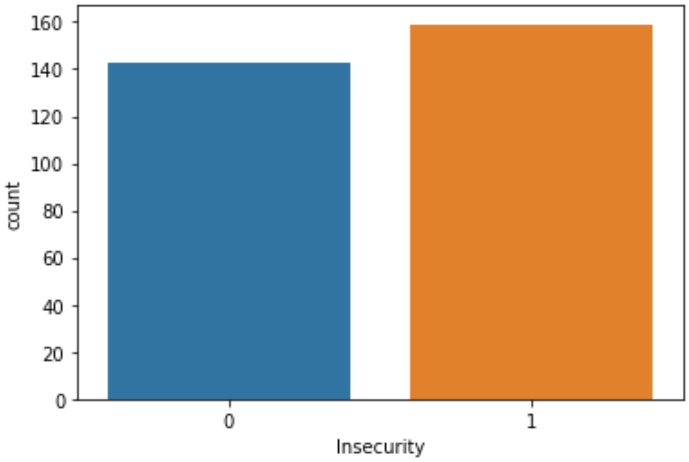


Fig.1.5

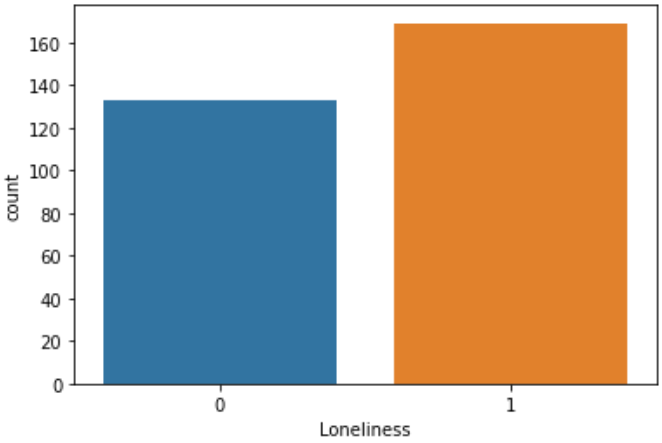


Fig.1.6

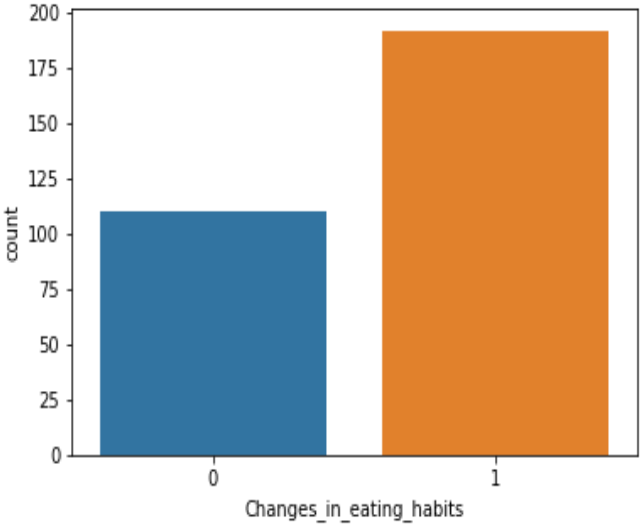


Fig.1.7

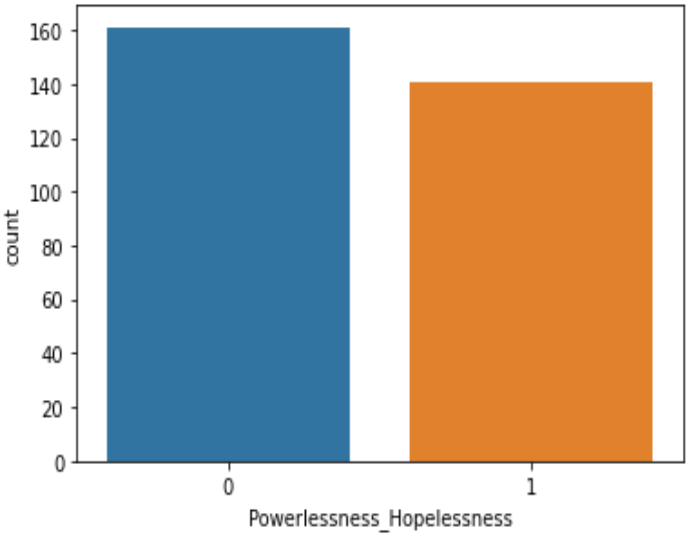


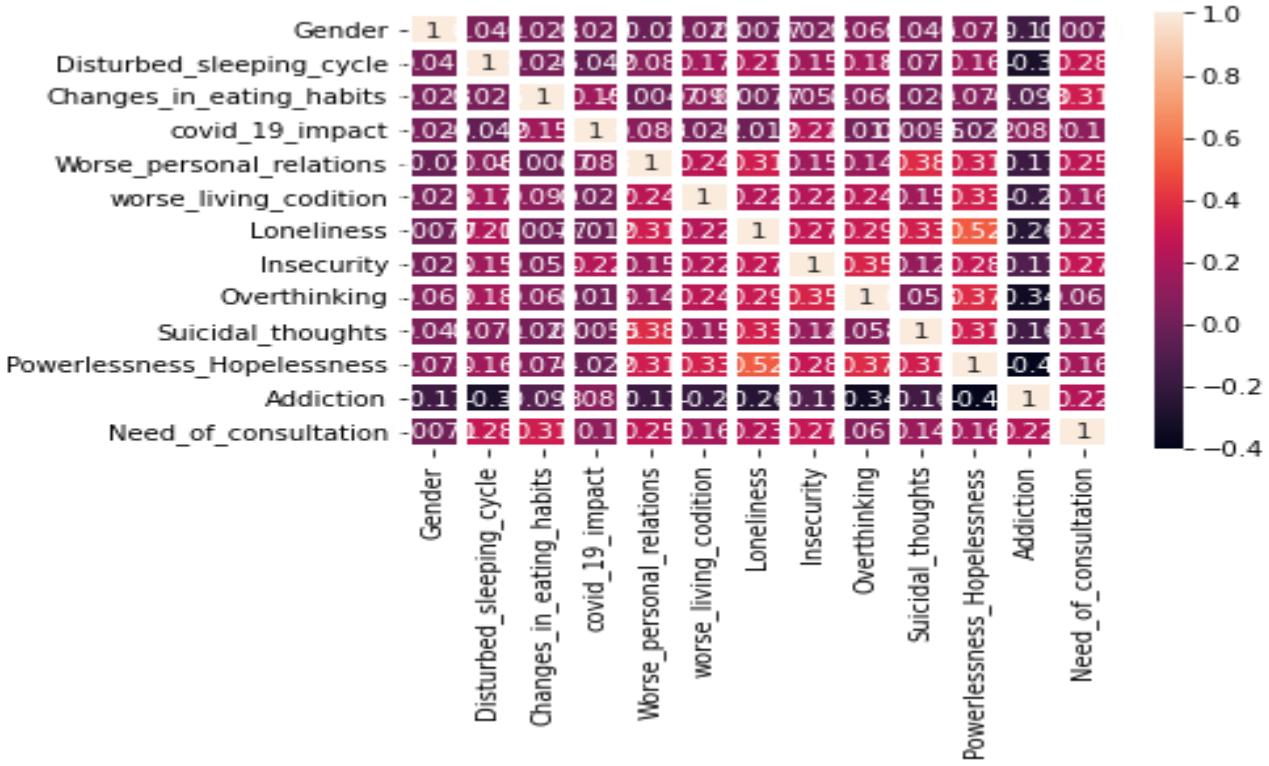
Fig.1.8

(Fig.1 Histograms for different variables in consideration)

Interpretations:

- From the Fig. 1.1, we can see that the no. of students who need medical counselling and consultation is larger than the no. of students who don't.
- In the Fig. 1.1, we can conclude that around **79%** students are in need of medical counselling and consultation.
- As shown in the the Fig. 1.1, the the percentage of female students who need medical counselling and consultation is larger than that of male percentage.
- We can see from the Fig. 1.2, around **66.2%** students have experienced consistent **disturbance in their sleeping cycle** due to pandemic.
- As observed in the Fig.1.3, around **66.2%** students **have not experienced worse personal relationships** in their lives due to pandemic.
- As seen in the Fig. 1.4, around **66.2%** students tend to **overthink** everything.
- We can observe from Fig. 1.5 that around **53% students experience insecurity about their future .**
- As shown in the Fig.1.6, around **53% students experience loneliness in their life.**
- From the Fig.1.7, we can observe that around **59.6%** students have experienced consistent **changes in eating habits** due to pandemic.
- In the Fig.1.8 we can see that around **46% students feel hopeless in their lives.**

- ***Correlation Heatmap:***



Interpretation:

- From the above heatmap we can observe that the explanatory variables **aren't highly correlated to each other**.

- **Model Building:**

I have used Random Forest Classifier, Decision Tree Classifier, KNN Classifier, Logistic Regression and Naïve Bayesian Classifier algorithms. These models have been selected for the study because of their popularity in recent literature. Before the model building by using chi-square test of independence, it was concluded that the explanatory variables are independent of each other.

1. Random Forest Classifier:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

2. Decision Tree Classifier:

Decision tree are powerful classification algorithm that are becoming increasingly popular with the growth of data mining in the field of information systems. In the literature there are so many popular decision tree algorithms like ID3, C4.5 and C5 etc. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms like information gain, Gini index to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is conducted. The objective of the splitting pair that maximizes the homogeneity of the resulting two or more subgroup samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain used in ID3, C4.5 and C5. Gini index is used in CART. In this J48 algorithm has been used as default by the software as it is a simple C4.5 decision tree for classification.

3. Naïve Bayesian Classifier:

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of class variable. This conditional independence assumption rarely holds true in real world application, hence the characterization as naïve yet the algorithm tends to perform well and learn rapidly in various supervised classification problems. Naïve Bayes classifier is based on Bayes theorem and the theorem of total probability.

In this classifier we compute the conditional probability P(Cj|X) and assign X to those class Ci having large probability i.e. X belongs to Cj if P(Cj|X)> P(Ci|X) for all

$i \neq j=1,2,\ldots ,m$ where m is no. of classes.

4. **Logistic Regression:**

Logistic regression is generalisation of linear regression. It is used primarily for predicting binary outcome. Because the response variable is discrete , it cannot be modelled linear regression. Therefore, rather than predicting point estimate of the event itself, it builds the model to predict the odds of its occurrence. In two class problem, odds greater than 50% would mean that the case is assigned to the class designed as "1" and "0" otherwise. While logistic regression is a powerful modelling tool, it assumes that the response variable is linear in the coefficients of the predictor variable.

❖ Comparison between different classification techniques:

A big part of machine learning is classification we want to know what class (a.k.a. group) an observation belongs to. The ability to precisely classify observations is extremely valuable for various applications. Data science provides a plethora of classification algorithms such as logistic regression, KNN, naive Bayes classifier, and decision trees, etc.

This project work presents a comparison and selection of different machine learning classification techniques applied to conclude if the student is in need of medical counselling and consultation or not using the collected data. The selected classifiers techniques can be applied to predict whether the student needs medical counselling and consultation.

- ## Measures for performance evaluation:

1. **Accuracy:** Accuracy is a ratio of correct predictions to the total predictions.

$$\text{Accuracy} = (TN+TP)/N$$

2. **AUC:** Area under the curve(AUC) in ROC curve provides a way to measure accuracy of a classification. The larger area gives more accurate classification.

3. **Recall**: What percent of the positive cases did you catch?

Percentage of correct positive predictions relative to total actual positives.

$$\text{Recall} = TP/(TP+FN)$$

4. **Precision**: What percent of your predictions were correct?
   Percentage of correct positive predictions relative to total positive predictions.
   Precision= TP/(TP+FP)

5. **F1 Score:** A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

   F1 Score: 2 * (Precision * Recall) / (Precision + Recall)

The following table represents the different classification techniques arranged in ascending ordered according to their accuracy.

Model	Accuracy	AUC	Recall	Precision	F1-Score
Random Forest Classifier	0.9235	0.9875	0.9929	0.9225	0.9549
Decision Tree Classifier	0.9055	0.8717	0.9352	0.9536	0.9408
K-Nearest Neighbours Classifier	0.8941	0.8952	0.9709	0.9068	0.9364
Logistic Regression	0.8585	0.9090	0.9566	0.8805	0.9155
Naive Bayes	0.8107	0.8668	0.8363	0.9109	0.8652

Interpretation:

From the above table, it is clear that accuracy of random forest classifier is larger than all oher classifiers(**92.35%**) while the precision of Decision tree classifier(**95.36%**) is greater than all other classifiers.

- **VISUALIZATON OF PERFORMANCE EVALUATION MEASURES:**

1. Confusion Matrix:

- A confusion matrix is a way to express how many of a classifier's predictions were correct, and when incorrect, where the classifier got confused (hence the name).
- In the confusion matrices below, the rows represent the true labels and the columns represent predicted labels. the column tells us what the classifier predicted, and the row tells us what the right label was. This is a convenient way to spot areas where the model may need a little extra training.To get even more insight into model performance, we examine other metrics like precision, recall, and F1 score.

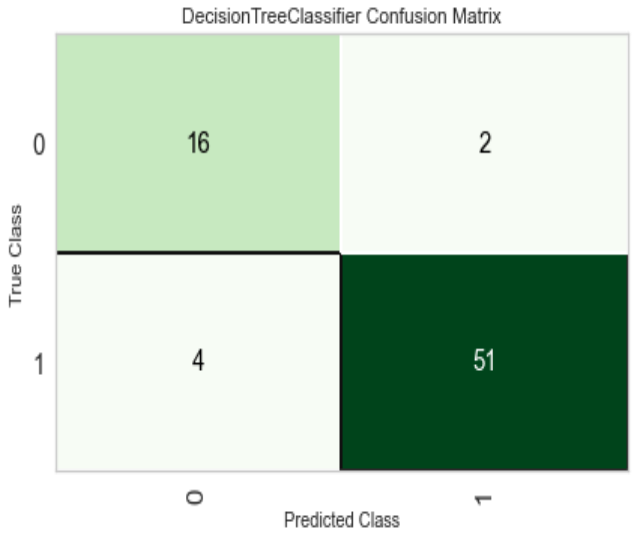


Fig.2.1 Decision tree classifier

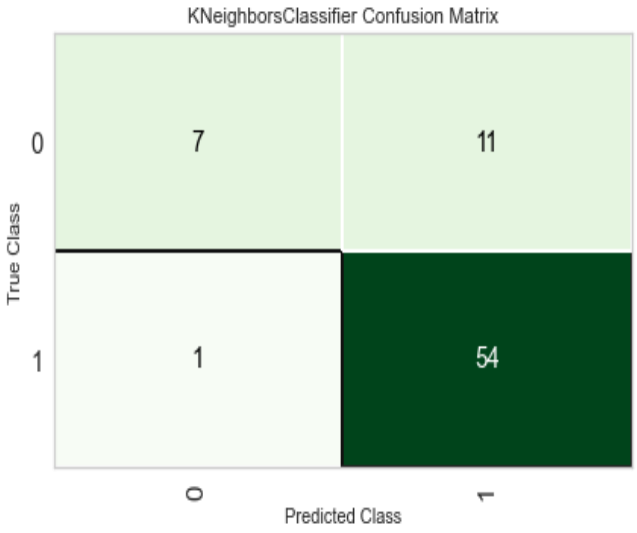


Fig. 2.2 K-nearest neighbours classifier

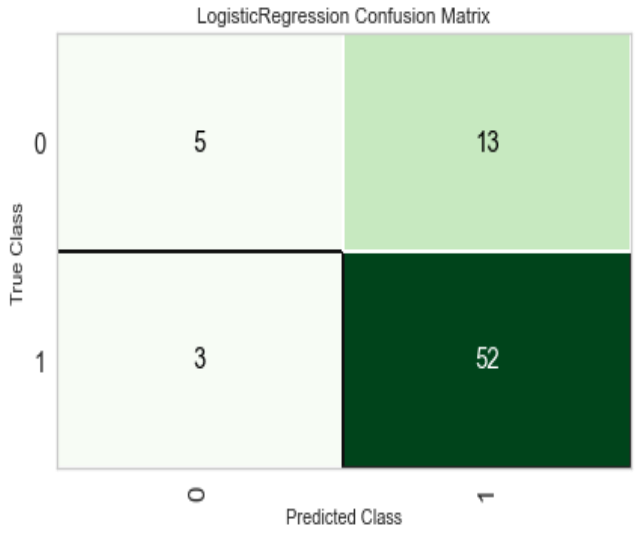


Fig. 2.3 Logistic regression classifier

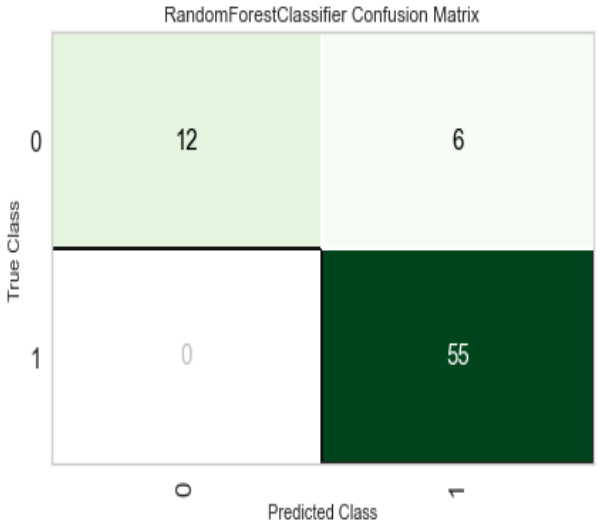


Fig. 2.4 Random forest classifier

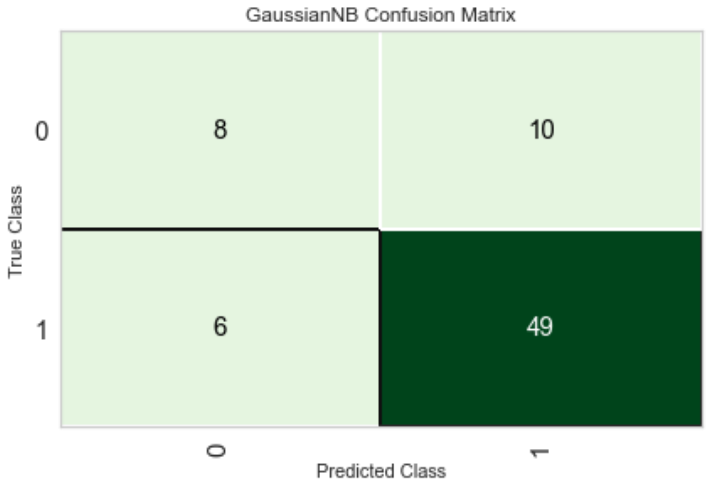


Fig. 2.5 Naïve Bayesian classifier

(Fig.2 Confusion matrices for different classifiers)

Interpretation:

- From above confusion matrices, we can conclude that the classifiers random forest and decision tree perform better than KNN, logistic regression and Naïve Bayes.

2. Classification Report :

- The classification report visualizer displays the precision, recall, F1, and support scores for the model.
- On its own, a classification report tells us generally what kind of errors the model made, but it doesn't give us specifics. The confusion matrix tells us exactly where mistakes were made, but it doesn't give us summary metrics like precision, recall, or F1 score. Using both of these can give us a much more nuanced understanding of how our model performs, going far beyond what an accuracy score can tell us and avoiding some of its pitfalls.

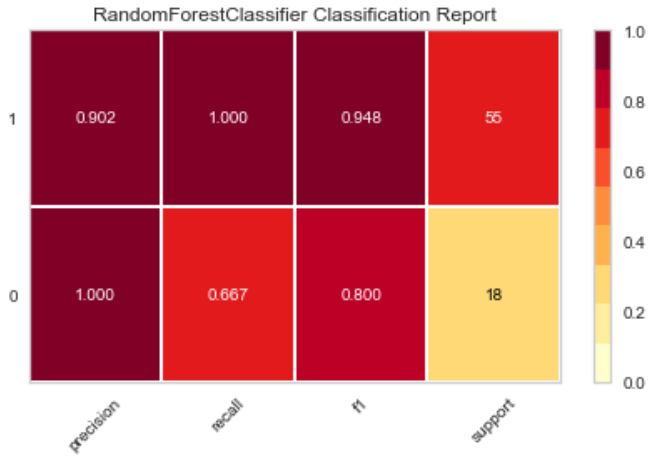


Fig. 3.1 Random forest classifier

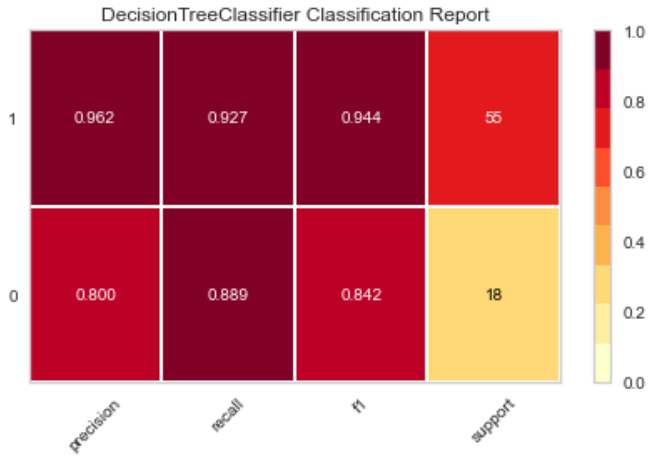


Fig.3.2 Decision tree classifier

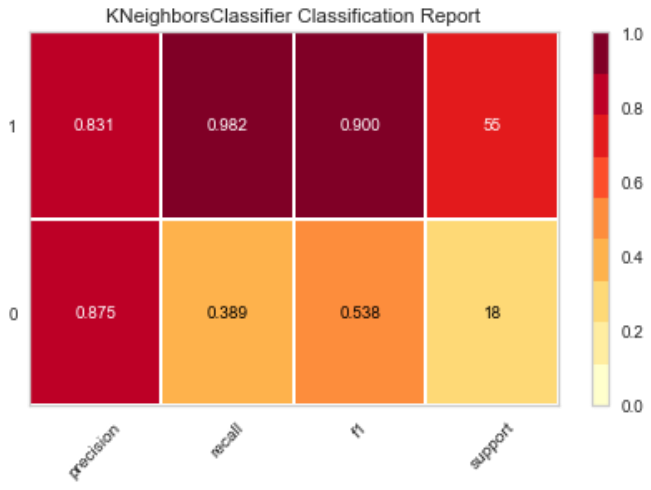


Fig. 3.3 K-nearest neighbours classifier

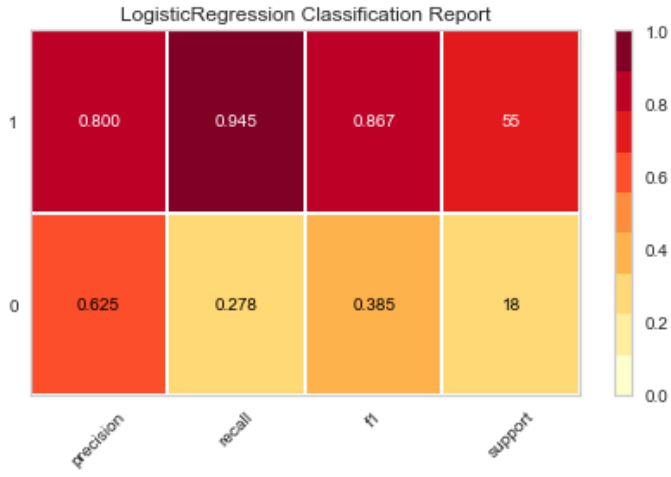


Fig. 3.4 Logistic regression classifier

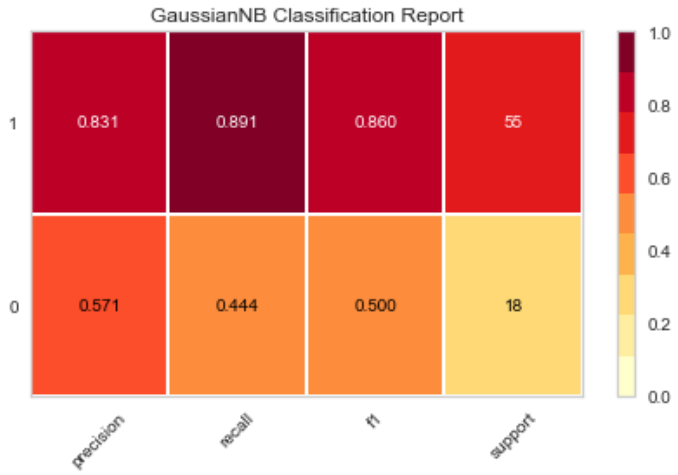


Fig. 3.5 Naïve Bayesian classifier

(Fig. 3 Classification report for different classifiers)

Interpretation:

- Check out the metrics for:

  Random forest classifier at level 0, KNN Classifier at level 0 and Logistic regression at level 0. **Precision is high**, meaning that the model was careful to avoid labelling students "No need of medical counselling and consultation" who don't really need it. On the other hand, **recall is relatively low**, which means that the classifier is missing a bunch of correct labels because it is being too careful! The F1 score reflects this imbalance.

3. Class Prediction Error:

- The Class Prediction Error plot is a twist on other and sometimes more familiar classification model diagnostic tools like the Confusion Matrix and Classification Report. Like the Classification Report, this plot shows the support (number of training samples) for each class in the fitted classification model as a stacked bar chart. Each bar is segmented to show the proportion of predictions (including false negatives and false positives, like a Confusion Matrix) for each class.
- This can often enables a better understanding of strengths and weaknesses of different models.

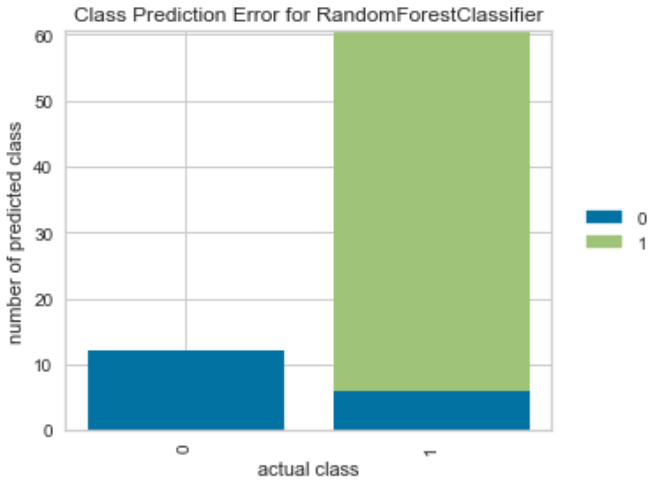


Fig. 4.1 Random forest classifier

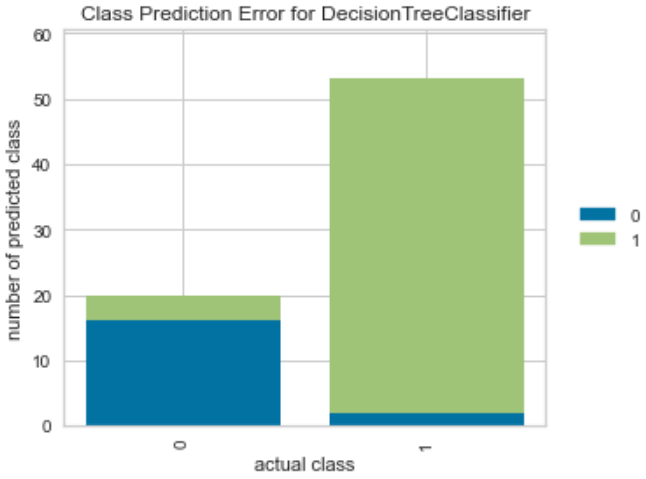


Fig.4.2 Decision tree classifier

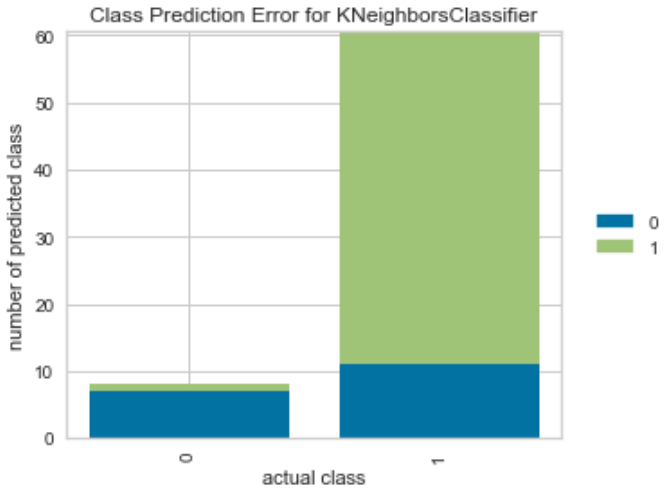


Fig. 4.3 K-nearest neighbours classifier

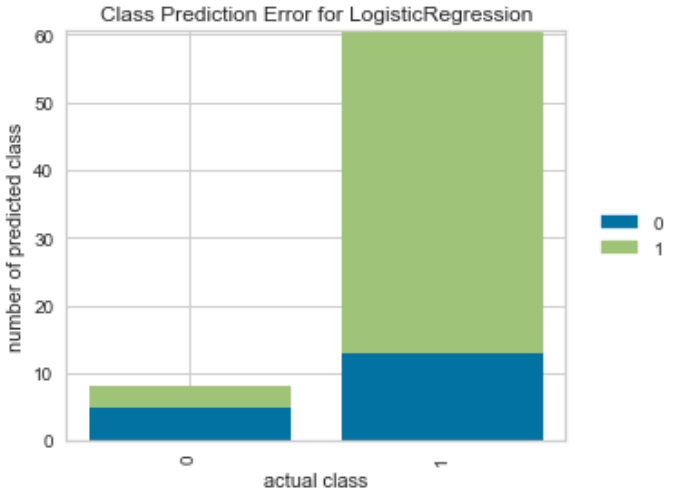


Fig. 4.4 Logistic regression classifier

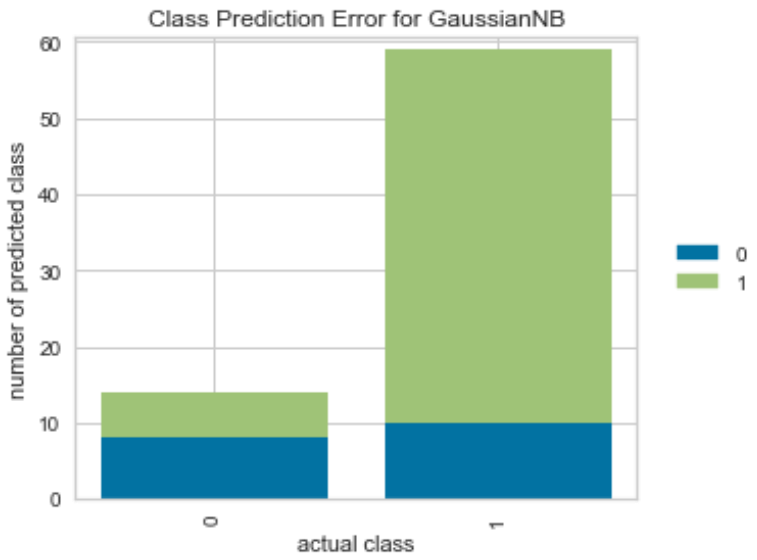


Fig. 4.5 Naïve Bayesian classifier

(Fig. 4 Class prediction error plot for different classifiers)

Interpretations:

- In the Fig 4.1, while the **Random Forest Classifier** appears to be fairly good at correctly predicting a class "No need of consultaion" based on the features of the students, it sometimes incorrectly class "Need of medical counselling and consultation".

- From the Fig 4.2, we can say that, while the **Decision Tree Classifier** appears to be fairly good at correctly predicting both the classes.

- As observed in Fig 4.3, while the **KNN Classifier** appears to be fairly good at correctly predicting a class "No need of consultaion" based on the features of the students, it often incorrectly class "Need of medical counselling and consultation".

- As shown in the Fig 4.4, while the **Logistic Regression Classifier** appears to be fairly good at correctly predicting a class "No need of consultaion" based on the features of the students, it often incorrectly class "Need of medical counselling and consultation".

- In the Fig. 4.5, while the **Naïve Bayes Classifier** appears to be least competing at correctly predicting both the classes compared to other models.

4. ROC Curve:

- In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of classification problem, we use the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model's performance.

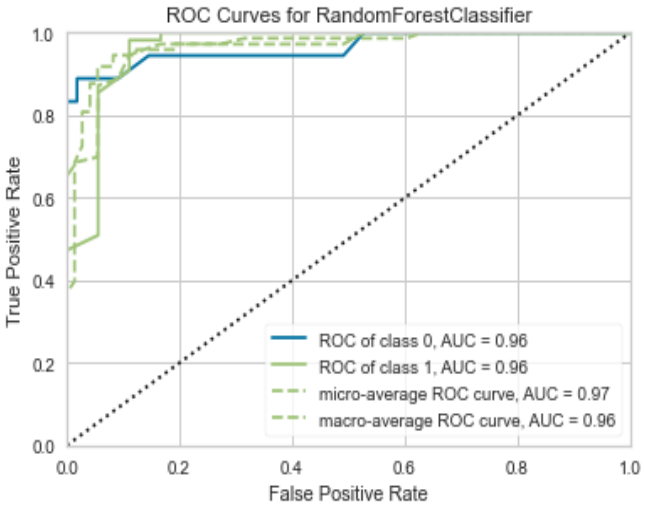


Fig. 3.1 Random forest classifier

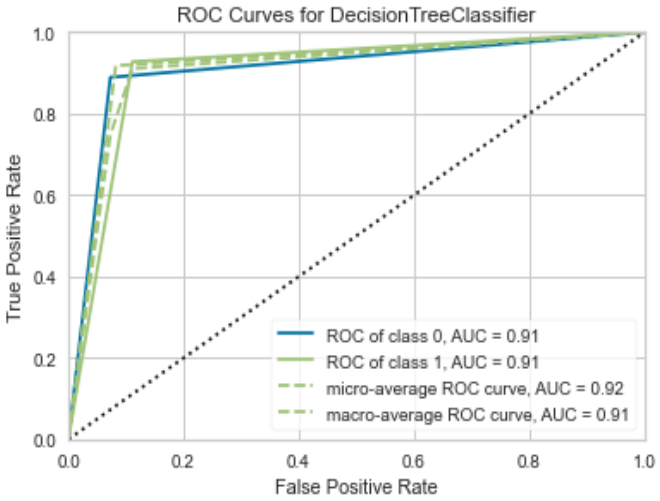


Fig.3.2 Decision tree classifier

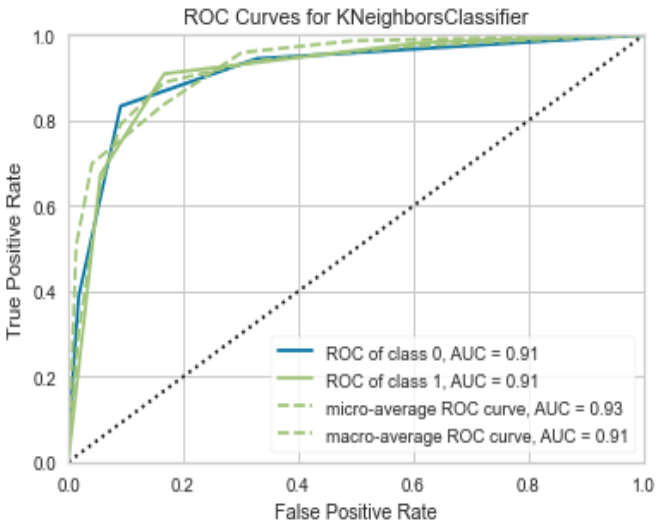


Fig. 5.3 K-nearest neighbours classifier

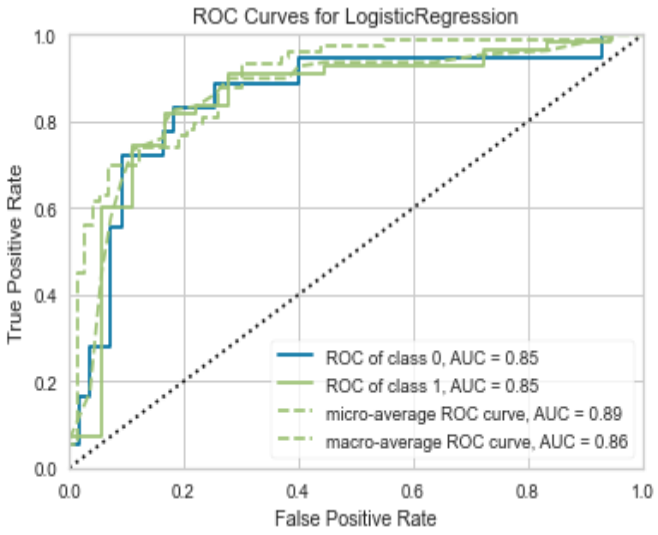


Fig. 5.4 Logistic regression classifier

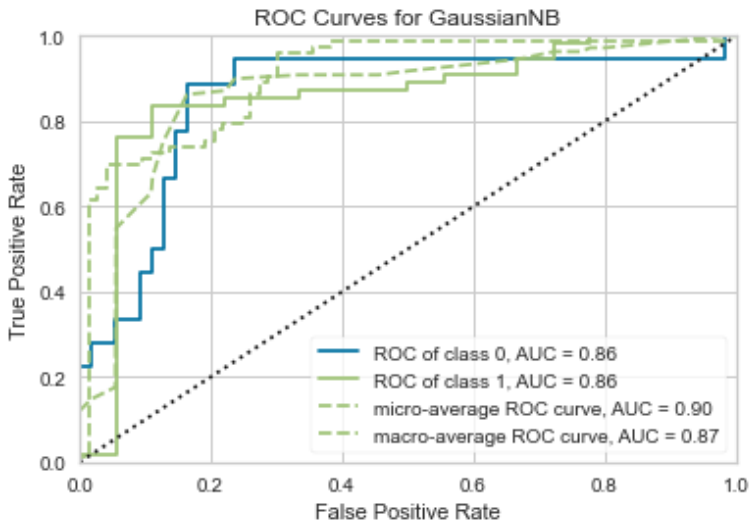


Fig. 5.5 Naïve Bayesian classifier

(Fig. 5 ROC Curves for different classifiers)

Interpretation:

From the above ROC curves, we can conclude that the random forest classifier has the best performance with (AUC=0.96) among all other classifiers.

5. Feature Importance Plot:

Feature Importance refers to techniques that calculate a score for all the input features for a given model the scores simply represent the "importance" of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable.

1. Random Forest Classifier:

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

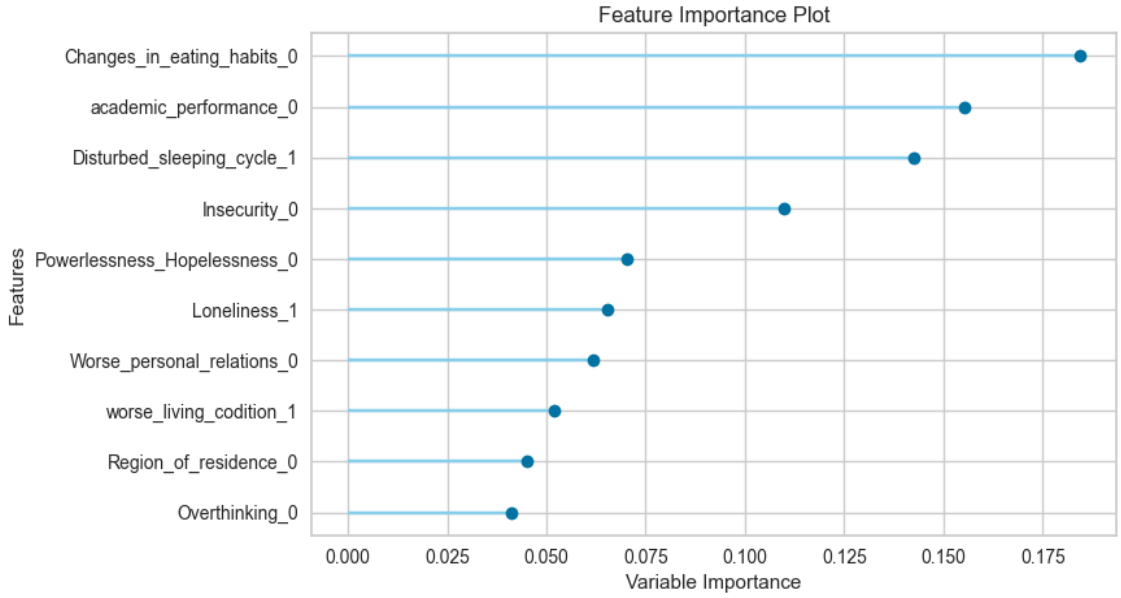


Fig. 6.1 Random forest classifier

Interpretation:

From above plot, we can conclude that the "Changes in eating habits" has the most effect on prediction of class " Need of medical counselling and consultation" where the feature "gender" is least effective.

2. Decision Tree Classifier:

The way we have find the important feature in Decision tree same technique is used to find the feature importance in Random Forest.

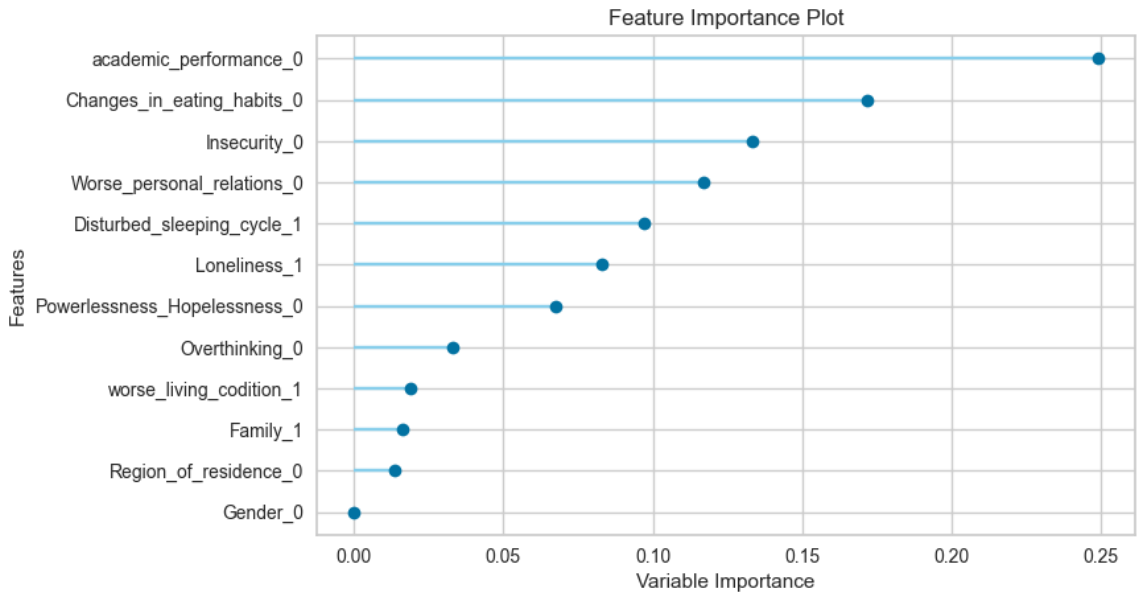


Fig. 6.2 Decision tree classifier

(Fig. 6 Feature importance plot)

Interpretation:

From above plot, we can conclude that the "Academic Performance" has the most effect on prediction of class " Need of medical counselling and consultation" where the feature "gender" is least effective.

Major Findings

1. Around **79%** students in Shivaji University are in need of medical counselling and consultation.

2. The the percentage of female students who need medical counselling and consultation is larger than that of male percentage.

3. Around **59.6%** students have experienced consistent **changes in eating habits** due to covid 19.

4. Around **66.2%** students have experienced consistent **disturbance in their sleeping cycle** due to covid 19.

5. Around **66.2%** students **have not experienced worse personal relationships** in their lives during covid 19.

6. Around **66.2%** students tend to **overthink** everything.

7. Around **53% students experience insecurity about their future .**

8. Around **53% students** experience **loneliness** in their life.

9. Around **46% students** feel **hopeless** in their lives.

10. Conclusions from model building:

We further modelled our data using various classification techniques such as random forest classifier, logistic regression, KNN, naive Bayes classifier, and decision tree classifier. We used Accuracy, ROC curve, precision, recall, etc. to compare the results. In conclusion, we can show that it is possible to use a superior analytics algorithm through the use of different classifiers. In terms of accuracy, random forest classifier is best than all other classifiers while in terms of precision, decision tree has it greater than all other classifiers

SCOPE AND LIMITATIONS

✓ Scope for the project:

- There is a scope to implement this project on lager polpulation.
- The project can also be extended to different parts of population such as buisnessmen, farmers, youth, elderly people,children, teenagers,etc.

✓ Limitations of the project:

- The analysed data is of particular university. So the tools applied to the data can change their performance if we change the data.
- The results of statistical tools comes from this particular analysis are not valid universally.
- We develop model only with available variables but if add another important variables then we expect that our models give better results.

Statistical tools and software used for analysis

- Graphical Representation
- Classification Techniques: Random Forest Classifier, Decision tree classifier, KNN classifier, Logistic regression classifier and Naïve Bayes classifier.
- MS-Excel
- Minitab
- Python

REFERENCES

1. Dr. Pankaj Kupwade (M.B.B.S.,D.C.H)

2. J. Han and M. Kamber, “Data Mining Concepts and Techniques”, Elevier, 2011.

3. Paul Ratanasiriong . Takashi China . Shiho Toyama : “ Mental health and well being of University students in Okinawa”

4. Ayesha Kamran Ul haq . Amira Khattak . Noreen Jamil . M. Asif Naeem . Farhaan Mirza: “ Data Analytics in Mental Health Care”

5. Xiaoli Zhang: “ Problems and countermeasures of college students’ mental health education”

6. Wes McKinney: “Python For Data Science”

7. www.pycaret.org

8. www.wikipedia.org

APPENDIX

- Link to the questionnaire: https://docs.google.com/forms/d/e/1FAIpQLSe9i2RcbL7H-A17QtOht4fu-5BR1NCnwLNa528o2rp0dkOAQg/viewform?usp=sf_link

- **Python code for model building:**

```
import pandas as pd
import numpy as np
import seaborn as sns
from pycaret.classification import *
import pandas as pd
#import the data from excel
data=pd.read_csv('C:/Users/DS_USER/Desktop/Project On Mental Health/final coded data for python.csv')
#dropping unnecessary columns
df=data.drop(['sr_no','Relationship_Status','Stream','Suicidal_thoughts','covid_19_impact','Addiction','Addiction_in_family'],axis=1)
#check if all data is imported
df.head()
#Information of data
df.info()
#EDA
sns.countplot(x='Need_of_consultation',data=data1)
sns.countplot(x='Need_of_consultation',hue='Gender',data=data1)
sns.countplot(x='Disturbed_sleeping_cycle',data=data1)
sns.countplot(x='Overthinking',data=data1)
sns.countplot(x='Insecurity',data=data1)
sns.countplot(x='Powerlessness_Hopelessness',data=data1)
sns.countplot(x='Loneliness',data=data1)
sns.countplot(x='Worse_personal_relations',data=data1)
sns.countplot(x='Changes_in_eating_habits',data=data1)
correlation=data1.corr()
correlation
sns.heatmap(correlation,xticklabels=correlation.columns,yticklabels=correlation.columns,annot=True,linewidths=3.5)
#splitting the data into seen and unseen data
```

```
train_data=df.sample(frac=0.8,random_state=786).reset_index(drop=True)
test_data=df.drop(train_data.index).reset_index(drop=True)
#setting up the experiment
experiment=setup(train_data,target='Need_of_medical counselling and
consultation',remove_multicollinearity=True,remove_perfect_collinearity=True
)
#comparing models
my_models=compare_models()
#prediction
predict_model(my_models,test_data.head())
rf=create_model('rf')
save_model(rf,model_name=' Random Forest Classifier')
#creating other models
dt=create_model('dt')
save_model(dt,model_name='Decision Tree Classifier')
#to load a already saved model
load_model('Decision Tree Classifier')
#to predict from specific model
dt.predict(test_data.head())
#evaluation of model
evaluate_model(dt)
rf=create_model('rf')
save_model(rf,model_name='Random Forest Classifier')
evaluate_model(rf)
rf.predict(test_data.head())
knn=create_model('knn')
save_model(knn,model_name='K Nearest Neighbour Classification')
evaluate_model(knn)
knn.predict(test_data.head())
knn.predict(train_data)
lr=create_model('lr')
save_model(lr,model_name='Logistic Regression')
evaluate_model(lr)
lr.predict(test_data.head())
nb=create_model('nb')
save_model(nb,model_name='Naive Bayes')
evaluate_model(nb)
nb.predict(test_data.head())
```