الرحيم الرحمن الله بسم

International Islamic University Chittagong

Department of Computer Science and Engineering

Course Title: Machine learning and data mining Lab

Course Code: CSE-4878

Session: Autumn-2023

Submission data: 22nd February 2023

Submitted by: -

Tasnimul Jannat Niha
C191242
C191242@ugrad.iiuc.ac.bd

Umme Kulsum Neha
C191272
C191272@ugrad.iiuc.ac.bd

Sadia Rahman Naima
C191264
C191264@ugrad.iiuc.ac.bd

8th semester
8BF

Submitted to: -

Mrs. Sabrina Akter

Department of CSE, IIUC

Introduction:

The dataset contains information about students' academic performance and their mental health. The objective of this study is to explore the relationship between the students' academic performance and their mental health status, and whether they have sought specialist treatment for mental health problems. The following variables are included in the dataset: timestamp, gender, age, course, year, CGPA, marital status, depression status, anxiety status, panic attack status, and whether they sought specialist treatment for mental health problems. The study will help in understanding the mental health status of the students and the impact it has on their academic performance.

Methods:

The study is a cross-sectional analysis of the dataset. The data was collected through an online survey, and the responses were recorded in a CSV file. Descriptive statistics were used to summarize the dataset. The dataset was cleaned, and any missing or inconsistent values were removed or replaced. The data was then analyzed using the programming language. The analysis included exploratory data analysis, and visualization of the data.

Dataset Characteristics with statistical model

Dataset: We have chosen the dataset of Student's Mental Health.

Dataset Information: This dataset has 11 features. The features are:

Timestamp	*Gender*	*Age*	*Course*	*Year*	*CGPA*	*Marital Status*	*Depression*	*Condition*	*Anxiety*	*Panic Attack*	*Specialist treatment (class)*

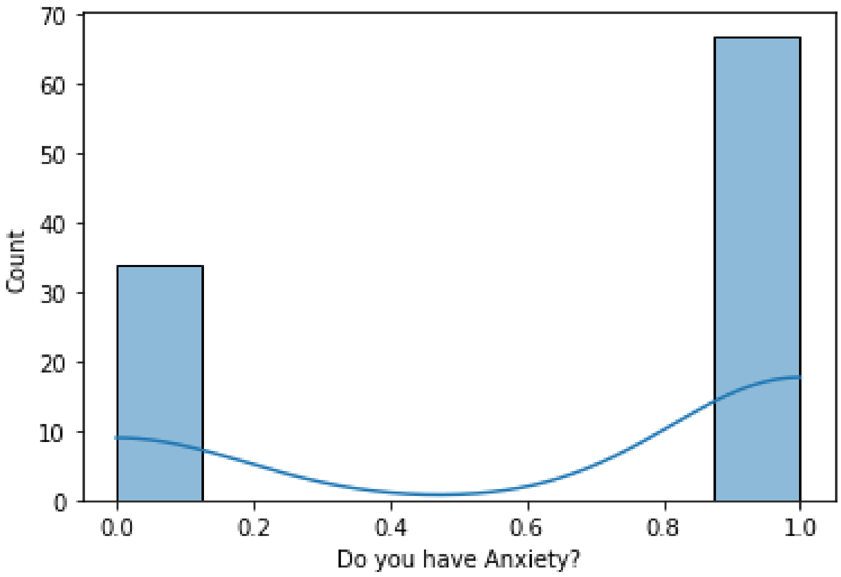
This dataset is collected from a survey in order to examine students' current academic situation and mental health.

Characteristics of the Dataset:

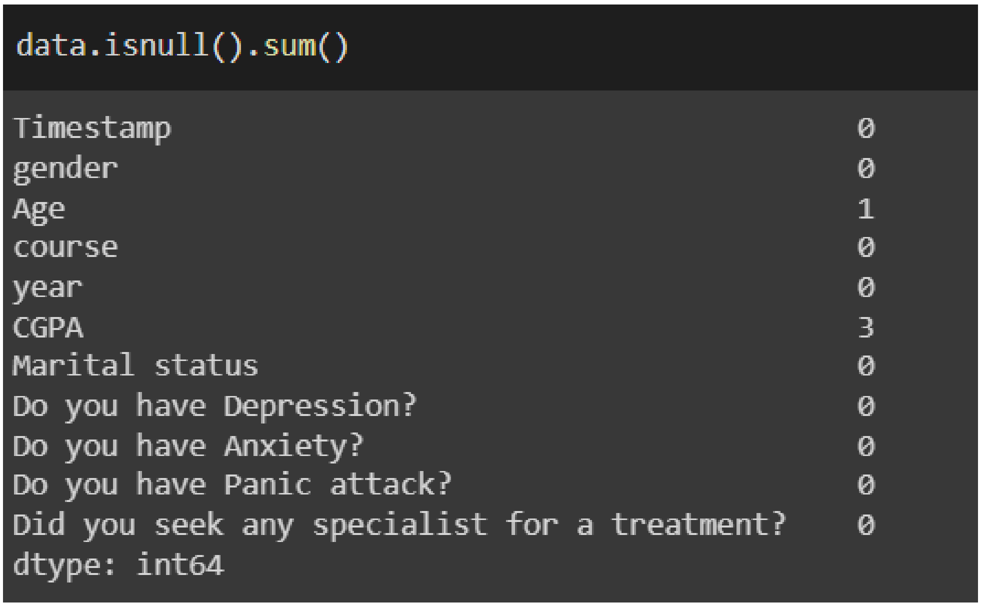
Dataset characteristics refer to the properties of a dataset, such as size, number of features, data types, and missing values. Understanding these characteristics is crucial in data analysis and modeling, as they can affect the accuracy and reliability of our models.

Normally distributed or not: In order to be considered a normal distribution, a dataset (when graphed) must follow a bell-shaped symmetric curve centered on the mean.

For Example, In this dataset, if we take the "Anxiety" attributes column of students we get mean of 2.53 and a standard deviation of 0.82. The histogram shows that the dataset is approximately normally distributed, with a peak around the mean value.



Find Missing Value: We found that there exists some missing value in our dataset. The "Age" attribute has 1 missing value & the "CGPA" attribute has 3 missing value.



```
data.isnull().sum()
```

```
Timestamp                                      0
gender                                         0
Age                                            1
course                                         0
year                                           0
CGPA                                           3
Marital status                                 0
Do you have Depression?                        0
Do you have Anxiety?                           0
Do you have Panic attack?                      0
Did you seek any specialist for a treatment?   0
dtype: int64
```

Filling Missing Value: As we know, for ordinal, interval and ratio data (if skewed) then we have to consider median value to consider the missing value. Thus we filled the missing value with Median value of those attributes.

Find Mean, Median & Mode: For filling the missing value of " Age" & "CGPA" column we need to find the mean, median & mode of these attributes.

	Mean	Median	Mode
Age	21.55	19.0	18.0
CGPA	3.495	3.49	4.0

Find Skewness: As we mentioned the mean, median & mode of Age & CGPA attributes, we can find the skewness through this information.

For "Age" attribute,

Mode < Median < Mean

18.0 < 19.0 < 21.55

So, this is positively skewed.

For "CGPA" attribute,

Mean < Median < Mode

3.495 < 3.49 < 4.0

Thus, it is negatively skewed.

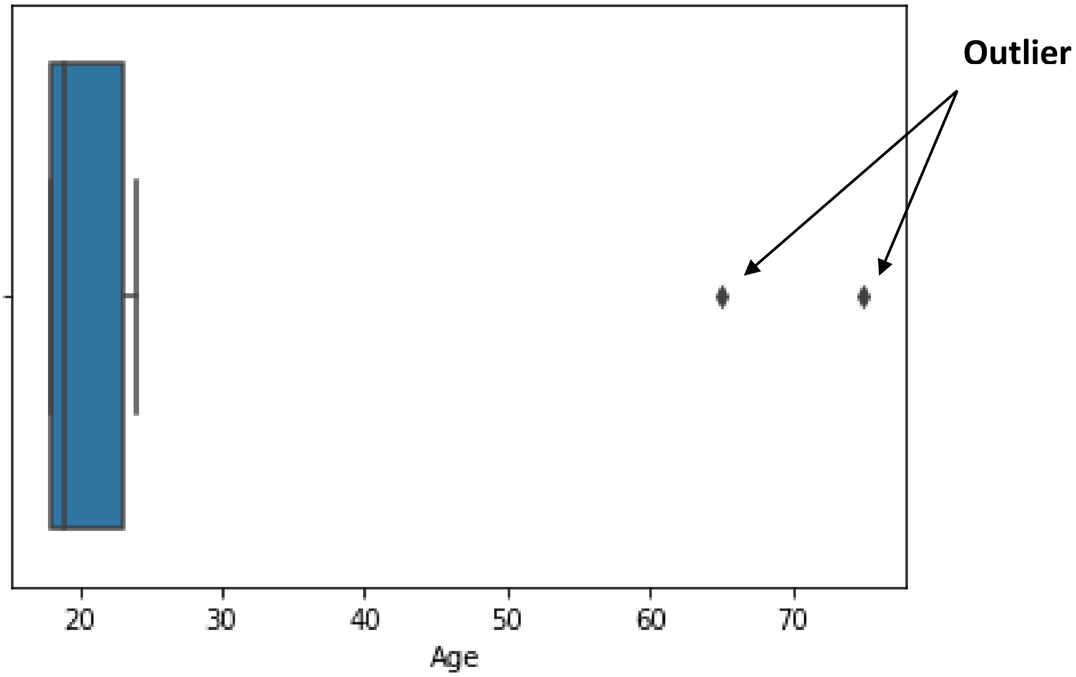
Find Variance: Variance is a measurement used to determine how far each number is from the mean and from every other number in the set.

	Age	CGPA
Variance	54.571881	0.285463

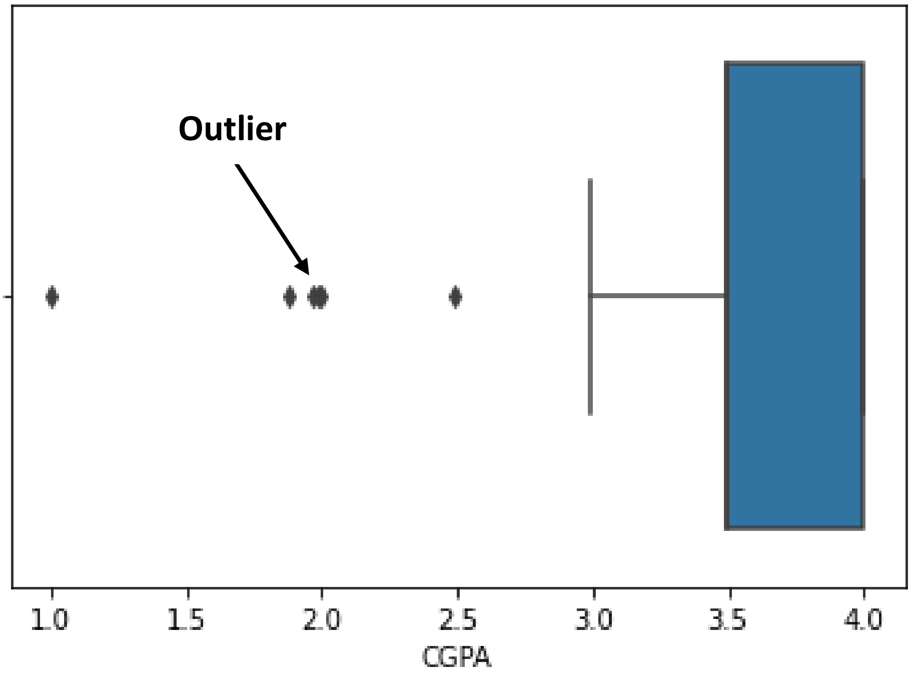
Detect Outlier: the outliers are the extreme values within the dataset. That means the outlier data points vary greatly from the expected values – either being much larger or significantly smaller.

For Detecting outlier we use the "Box Plot" method.

Outlier for Age attributes:



Outlier for CGPA attributes:

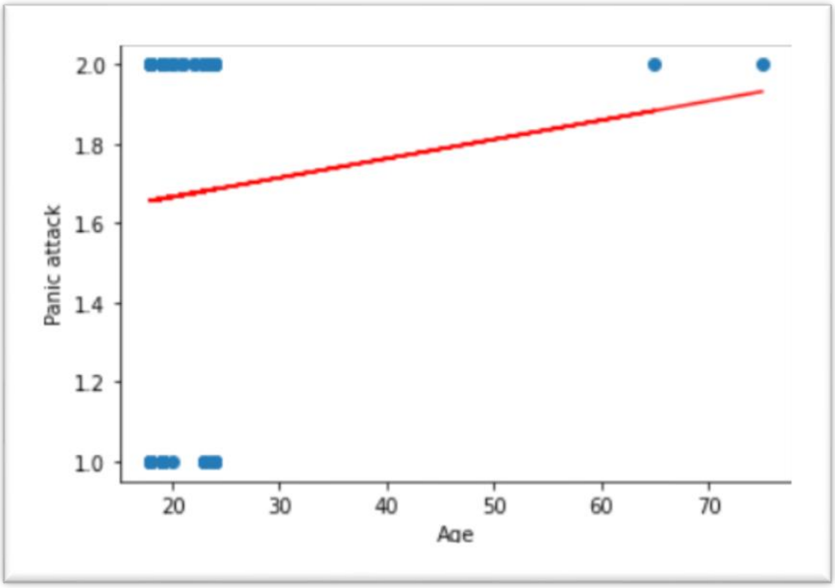


In summary, the dataset that we have chosen for this study is the Student's Mental Health dataset. It contains information on 11 features, including the timestamp, gender, age, course, year, CGPA, marital status, depression status, anxiety status, panic attack status, and whether the students sought specialist treatment for mental health problems. The dataset was collected through an online survey to examine students' current academic situation and mental health. In terms of dataset characteristics, we found that the data was normally distributed for some features, and there were missing values in the "Age" and "CGPA" attributes. We filled the missing values using the median values of those attributes, which were 19.0 and 3.49, respectively. The age attribute was positively skewed, while the CGPA attribute was negatively skewed. We also calculated the variance for each attribute, which was 54.57 for age and 0.29 for CGPA. Finally, we used a box plot to detect outliers within the dataset.

Linear Regression implementation

The linear regression model is then created using scikit-learn.

In the first example, the dependent variable is "Panic attack," and the independent variable is "Age." The fit function is used to fit the model to the data. The predict function is used to predict the value of the dependent variable for a given independent variable. The intercept and coefficient values are then calculated. A scatter plot is then created to visualize the data, and a line of best fit is plotted on the graph.

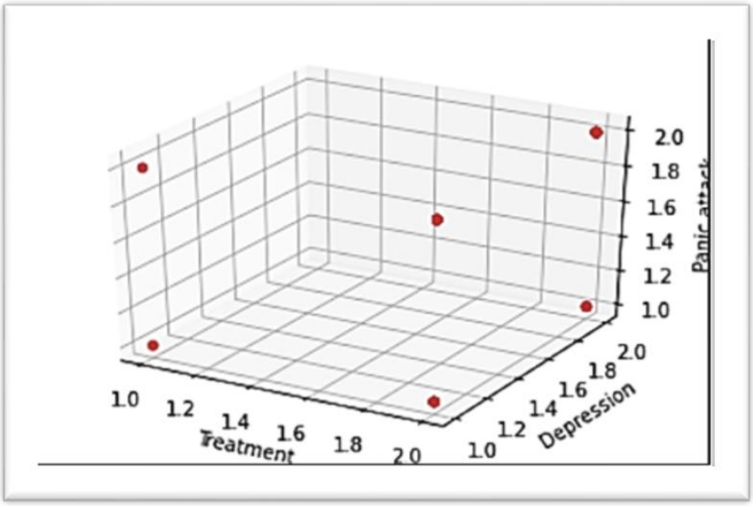


In the second example, the dependent variable is "Treatment," and the independent variables include "Age," "CGPA," "Marital status Modify," "Depression," "Panic attack," "Anxiety," and "Gender." The fit function is used to fit the model to the data. The predict function is used to predict the value of the dependent variable for a given set of independent variables. The intercept and coefficient values are then calculated.

Coefficients:

0.00035783	-0.02882739	0.13827039	0.10092874	0.03791948	-0.00348271	-0.0024848

In the third example, a 3D scatter plot is created to visualize the relationship between the dependent variable "Treatment" and the independent variables "Depression" and "Panic attack."



The linear regression model predicts

dependent variable	**Independent variable**	**R2 score**
Panic attack	Age	0.5712113152866571(single)
Treatment	ALL	16.051139319189044

Logistic Regression implementation

The dataset used in this code includes information about the students' demographics, academic performance, and mental health conditions. Logistic regression is a common statistical method used to predict binary outcomes. In the code, logistic regression is used to predict whether a student sought specialist treatment for depression, anxiety, or panic attacks based on the given variables.

First, the dataset is loaded and preprocessed. The categorical variables are transformed into dummy variables for use in the logistic regression model. Then, the model is trained using 70% of the data and tested on the remaining 30%. The accuracy of the model is evaluated using the confusion matrix and classification report.

The logistic regression model achieved an accuracy of **92.30769230769%** in predicting whether a student sought specialist treatment for mental health conditions. The confusion matrix shows that the model correctly predicted 74 cases and incorrectly predicted 15 cases. The classification report shows that the model has high precision, recall, and F1 score for predicting the positive class (students who sought treatment) but low values for the negative class (students who did not seek treatment). This may be due to the imbalanced nature of the dataset, with a majority of students not seeking treatment. Overall, the logistic regression model shows promise in predicting whether a student sought specialist treatment for mental health conditions.

Classifier regression	94.66666666666%
Accuracy	92.30769230769%

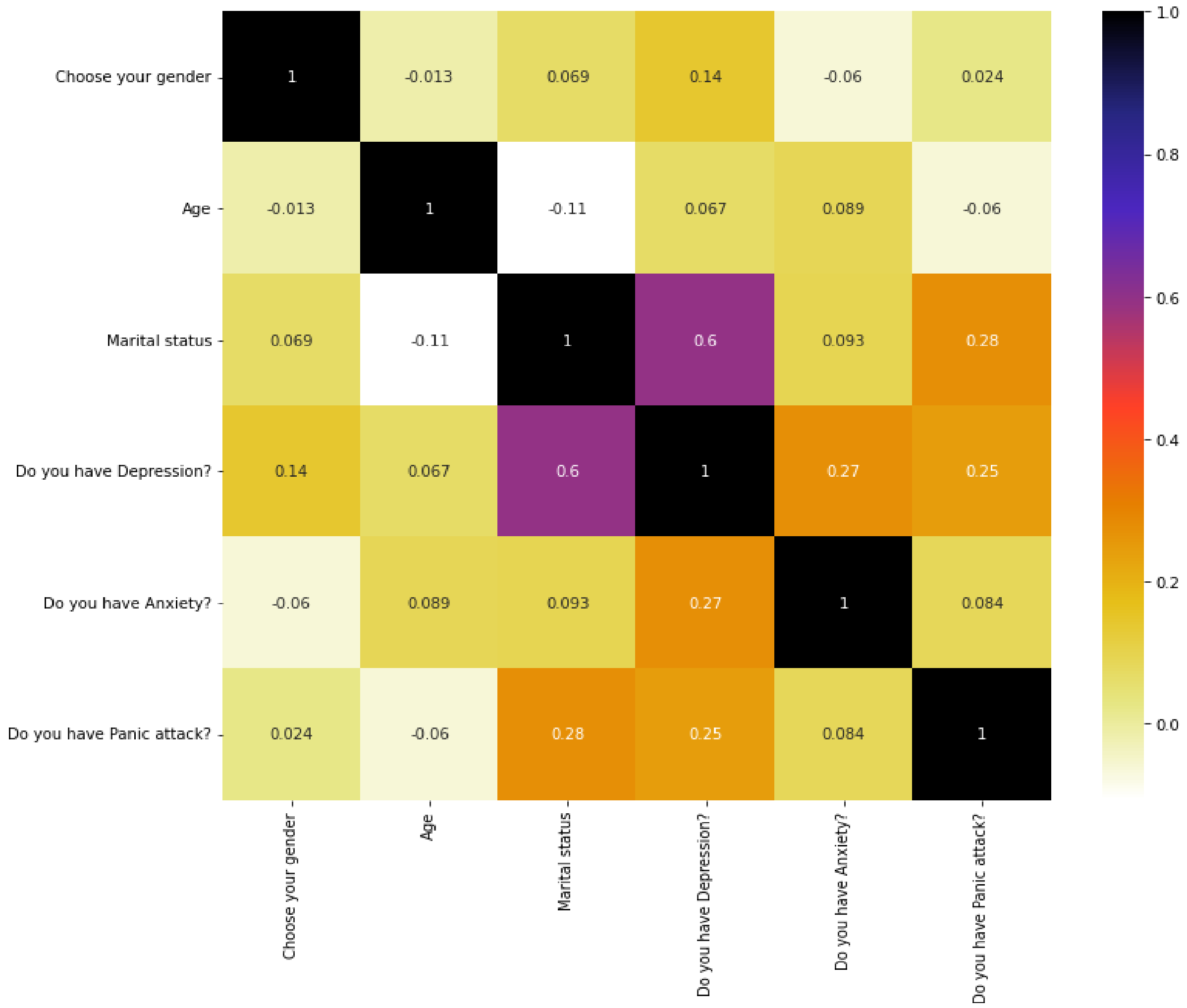
Feature selection

Pearson correlation: It is a statistical measure of the strength of a linear relationship between two variables. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

We take dataset on “Student’s Mental Health”. Now, let's use these statistical tests on the "student_mental_health.csv" dataset. This dataset contains data about the mental health of students, including various attributes such as age, gender, academic performance, etc.

Let's use the Pearson correlation to find the correlation between two continuous variables for all the Attributes to Attributes and Attributes to Class.

Correlation between Attribute to Attribute: All our independent attributes are Gender, Age, Marital status, Anxiety, Panic attack & Depression.



From the above plot of Pearson Correlation we can see that there consists both positive and negative correlation. The Pearson correlation coefficient between ‘Gender’ and ‘Age’ is -0.013, this means that when one variable increases as the other decreases, and vice versa. Most of the correlation lies between 0.06 – 0.09, which indicating a weak positive correlation between these attributes like ‘Anxiety’ & ‘Marital status’, ‘Age’ & ‘Depression’, ‘Anxiety’ & ‘Age’ etc. This means that as one attribute increases, there is a slight increase in other attributes level. We find that the most highly Correlation exists between the attribute “Depression” & “Marital Status” which is 0.6 means 60%.

Correlation between Attribute to Class: In our dataset our class is “Seeking specialist treatment”. We conclude a decision regarding this by analyzing our entire independent attribute.

	Seeking Specialist Treatment
Gender	0.052
Age	-0.049
Marital Status	0.35
Depression	0.35
Anxiety	0.087
Panic attack	0.18

From the above chart we can see that Class is highly correlated with two attributes "Marital status" & "Depression". It's also having weak correlation with other attributes.

Selection of Correlated Features: For finding the correlated features, Pearson correlation use Threshold values. Depending on this threshold the correlated features will selected. This threshold values is selected by us. Most of the time it will say that this Threshold value is at least 70% correlated, then we will remove the feature. But in our dataset as most of our attribute is object type & thus we converted this object data to numeric data for finding the pearson correlation. For this, most of our coefficient value lies 0.2 – 0.09 that means we can't even get 60-90% coefficient value for this data conversion. For this reason we select our Threshold value 0.2 as we get highest correlated features in this value.

Our Correlated features are –

Corr_features	Anxiety
	Depression
	Panic Attack

Chi-square test: It is a statistical test used to determine whether there is a significant association between two categorical variables. The test determines if the observed frequency of a category is significantly different from the expected frequency. Let's use the Chi-square test to find the correlation between two continuous variables for all the Attributes to Attributes and Attributes to Class.

F-value & P-value for feature selection: In the context of the chi-square test, the F-value and the p-value are two important statistical measures used to determine whether the null hypothesis should be rejected or not.

The F-value, also known as the test statistic, measures the degree of association between the two categorical variables. The p-value is a probability value that is used to determine whether the observed association between the two variables is statistically significant or just due to chance. A p-value less than the significance level (usually 0.05) indicates that the observed association is statistically significant, and we reject the null hypothesis. On the other hand, a p-value greater than the significance level suggests that the observed association is not statistically significant, and we fail to reject the null hypothesis.

In our dataset our obtaining F-value and p-values are –

Attribute	F-value	P-value
Gender	0.20411606	0.651419
Age	0.07318634	0.786753
Marital Status	1.9579773	0.161730
Depression	4.16842105	0.041184
Anxiety	0.25663786	0.612439
Panic attack	1.09484004	0.295401

The F-score needs to be higher. The more the F-score is the more important the feature is. & the P-value needs to be lower. The more lower the P-value is the more important that feature is. From the above table we can see that the attribute "Depression" has the larger F-value & the lowest p-value. Thus this feature is mostly important.

Selection of Correlated Features: From the chi square test we got below highly correlated features –

Corr_features	P_Values
Age	0.786753
Gender	0.651419
Anxiety	0.612439
Depression	0.041184
Panic Attack	0.295401
Marital Status	0.161730

In Chi square test we got the features that we also got in Pearson correlation. But in this test we got some extra features also that we didn't get in Pearson. This is because in Pearson we select the Threshold value by ourselves and we also select low value as we got maximum feature in that case. We choose the low value because of our data conversion. But

in Chi square test we need not have to take any value by ourselves. Thus we got some extra features along with old features.

PCA:

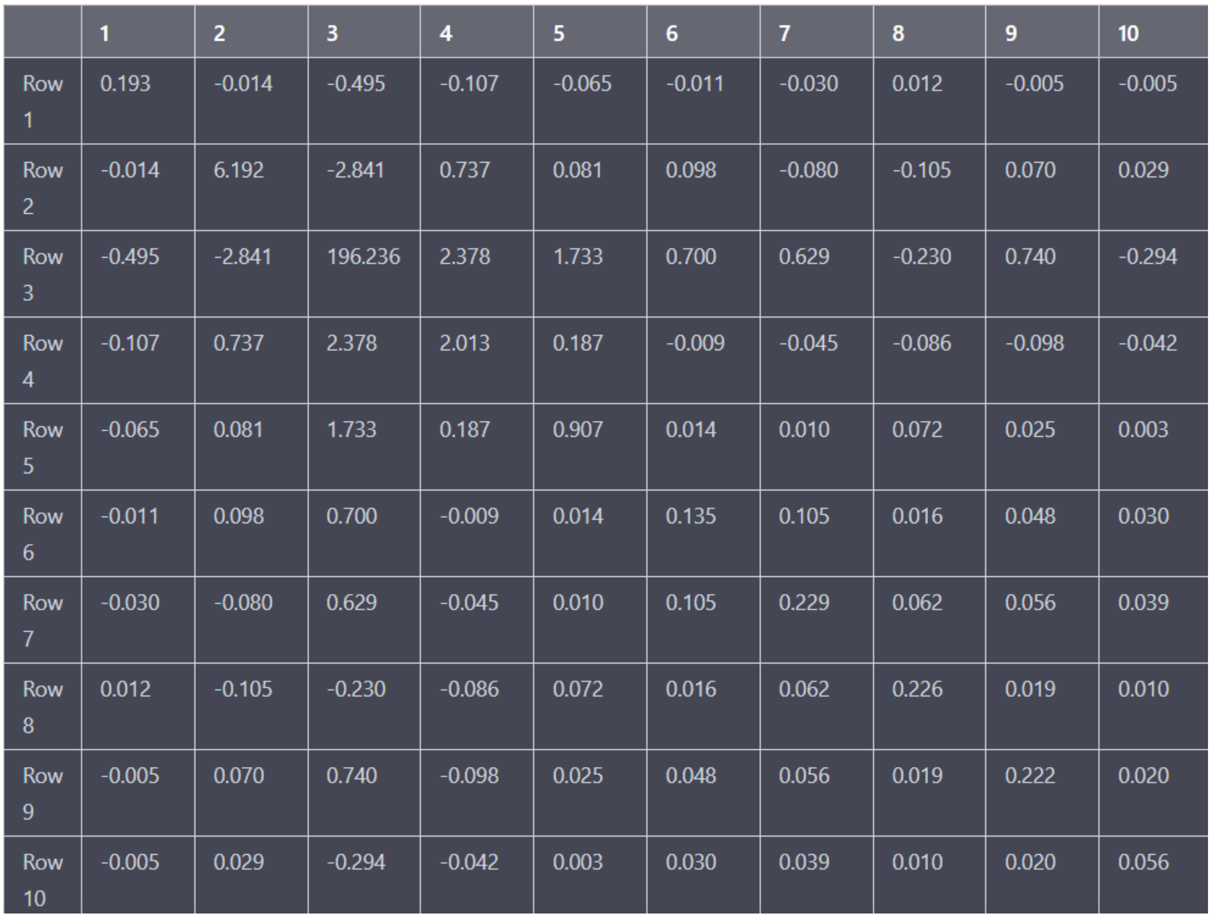
The objective was to use principal component analysis (PCA) to reduce the dimensionality of the chosen dataset.

Data Preprocessing:

- The "Timestamp" column was dropped from the dataset.
- The integer data type for the "Age" column was converted to float.
- The object data types for several columns (gender, course, year of study, CGPA, marital status, depression, anxiety, panic attack, and treatment history) were converted to float using label encoding.
- The dataset was then standardized using the StandardScaler function.

Covariance Matrix Analysis:

- A covariance matrix was generated using the numpy "cov" function.



	1	2	3	4	5	6	7	8	9	10
Row 1	0.193	-0.014	-0.495	-0.107	-0.065	-0.011	-0.030	0.012	-0.005	-0.005
Row 2	-0.014	6.192	-2.841	0.737	0.081	0.098	-0.080	-0.105	0.070	0.029
Row 3	-0.495	-2.841	196.236	2.378	1.733	0.700	0.629	-0.230	0.740	-0.294
Row 4	-0.107	0.737	2.378	2.013	0.187	-0.009	-0.045	-0.086	-0.098	-0.042
Row 5	-0.065	0.081	1.733	0.187	0.907	0.014	0.010	0.072	0.025	0.003
Row 6	-0.011	0.098	0.700	-0.009	0.014	0.135	0.105	0.016	0.048	0.030
Row 7	-0.030	-0.080	0.629	-0.045	0.010	0.105	0.229	0.062	0.056	0.039
Row 8	0.012	-0.105	-0.230	-0.086	0.072	0.016	0.062	0.226	0.019	0.010
Row 9	-0.005	0.070	0.740	-0.098	0.025	0.048	0.056	0.019	0.222	0.020
Row 10	-0.005	0.029	-0.294	-0.042	0.003	0.030	0.039	0.010	0.020	0.056

- The eigenvalues and eigenvectors of the covariance matrix were calculated using the numpy "eig" function.
- The eigenvectors were printed to the console.

Eigenvalues

1.96332358e+02	**6.29776521e+00**	**1.88628390e+00**	**8.842463 11e-01**	**3.596028 64e-01**	**2.102987 56e-01**	**1.831988 55e-01**	**1.506639 12e-01**	**6.105813 16e-02**	**4.383071 76e-02**

Eigenvectors

[[2.53467660e-03 6.59481691e-03 -5.96239727e-02 6.55868240e-02 -1.39060667e-01 5.36242184e-01 5.84151173e-01 5.79503380e-01 8.72354397e-02 2.30890333e-02]
[1.48838855e-02 -9.83263705e-01 -1.78121778e-01 -1.72560762e-03 -8.75417514e-03 1.74346380e-02 -2.23349748e-02 -5.41623753e-03 1.81782698e-02 -1.17121067e-03]
[-9.99751997e-01 -1.21665311e-02 -1.54890929e-02 7.42417464e-03 -5.09452256e-03 3.89008856e-03 -2.13433201e-03 -7.75107983e-04 5.39966914e-04 2.52082686e-03]
[-1.21846762e-02 -1.77354121e-01 9.67604622e-01 1.35734498e-01 8.12080005e-02 4.18416913e-02 7.12198160e-02 -4.95230114e-03 -3.00075502e-03 1.62244289e-02]
[-8.87127820e-03 -2.46737347e-02 1.41933154e-01 -9.76271500e-01 -1.12859503e-01 -3.99647343e-02 3.04537821e-02 1.01458816e-01 2.20881435e-02 -4.21754765e-03]
[-3.56455767e-03 -1.65786021e-02 -2.40939607e-02 -2.76253245e-02 4.29854259e-01 -1.09181581e-01 -3.26095430e-02 3.64699776e-01 -7.69847871e-01 -2.73748810e-01]
[-3.21416467e-03 1.27803774e-02 -2.73330160e-02 -4.23518556e-02 6.95036791e-01 -1.97080782e-02 -2.39497489e-01 3.42583819e-01 5.80729601e-01 -4.30281191e-02]
[1.16622061e-03 1.97608859e-02 -3.33031706e-02 -1.32789895e-01 3.31441460e-01 7.74968101e-01 -1.18096509e-01 -4.85424342e-01 -1.44406757e-01 -4.95197091e-05]
[-3.76580978e-03 -1.00495855e-02 -7.10510918e-02 -5.95351739e-02 3.95478267e-01 -3.05087032e-01 7.61320328e-01 -3.98167237e-01 5.59948042e-02 -2.30166346e-02]
[1.49840816e-03 -2.83146386e-03 -2.42694323e-02 -1.93893160e-02 1.64682142e-01 -5.30531220e-02 -1.69248047e-02 9.63532062e-02 -1.94064430e-01 9.60134653e-01]]

PCA implement:

- PCA was performed using the sklearn "PCA" function with 2 components.
- The scaled data was transformed using the PCA function.
- The dimensions of the scaled data and transformed data were printed to the console.
- The principal components of the dataset were obtained using the pca.components_ attribute.
- Before data preprocessing, the scaled data had a shape of (101, 9) and the transformed data had a shape of (101, 2).
- After data preprocessing, the scaled data had a shape of (101, 9) and the transformed data had a shape of (101, 2).
- The principal components of the dataset were the same before and after data preprocessing.

The performance of PCA in reducing the dimensionality of the dataset was the same before and after data preprocessing. However, data preprocessing is still an important step to ensure that the data is standardized and that the algorithm is more efficient. The use of PCA can help to identify the most important variables in the dataset and reduce the complexity of the data.

PCA components

-0.11705839, 0.00998172, 0.10994864, -0.11327506, 0.09786618, 0.53546516, 0.55689293, 0.24473495, 0.35747767, 0.40885426
0.44969334, -0.29795148, -0.35237349, -0.6083892, -0.39146624, -0.09132801, 0.00705599, 0.19125642, 0.03062326, 0.12467202

Drive link: https://drive.google.com/drive/folders/1dzpoO2nU-HgT1UpNvpsuVLdek7_naER8?usp=sharing

THANK YOU