MULTIMEDIA UNIVERSITY

TDS 2101 FUNDAMENTAL DATA SCIENCE
Individual Project

Lecturer: Madam Syuhaidah Azni
Group: TT1L

Topic: Predicting Mental Health Risk Factors Among College Students: A Comparative Analysis of Different Classification Algorithms

Student Name	Student ID
Irfan Daniel Teng Bin Mohd Taufiq Teng	1191201519

1. Introduction

Mental health is a critical issue that affects individuals of all ages, and it has become increasingly important in recent years to understand the factors that contribute to mental health among young adults. University students, in particular, are at a high risk for mental health issues such as depression, anxiety, and panic attacks. Many university students experience mental health concerns at various stages of their education. This becomes even more important when they near the end of their studies and consider their future options.With the increasing pressure to succeed academically and socially, university students are more susceptible to mental health problems than ever before.

Understanding the risk factors that contribute to mental health issues among university students is crucial to helping them get the support they need. In this study, we aim to use machine learning techniques to predict mental health risk factors among university students by analysing their demographic, academic and self-reported mental health symptoms data. We will compare the performance of different classification algorithms and identify the most important predictors of mental health risk among university students.

The fundamental goal of this research is to investigate the state of mental well-being in university students by utilising numerous personality factors and fields of specialty.

2. Problem Statement

The problem statement that I am going to study is the effects of students' gender, age, course, course year, CGPA and marital status on their mental health and to use machine learning techniques to predict mental health risk factors.

3. Motivation

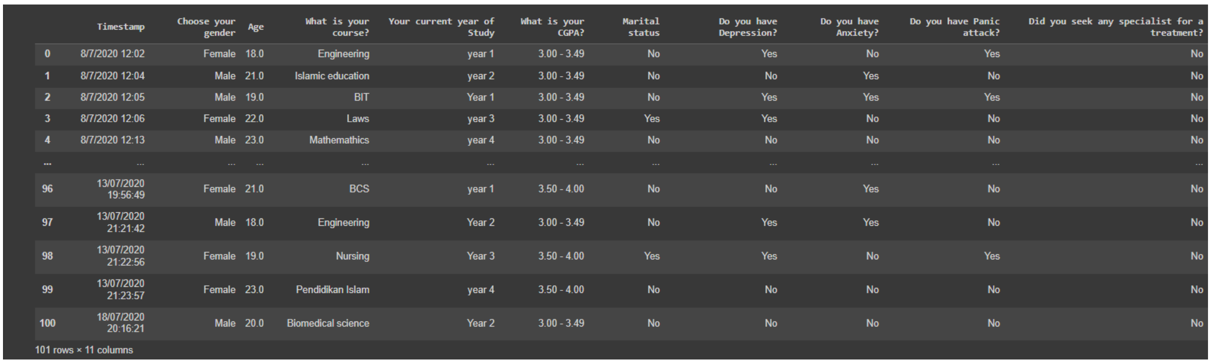
The motivation behind this study is to better understand the factors that contribute to mental health issues among university students and to use this knowledge to help support students in need. By using machine learning techniques to analyse demographic, academic, and self-reported mental health data, we aim to predict mental health risk factors among university students. We will compare the performance of different classification algorithms to understand which one performs the best in identifying mental health risk among university students. By identifying the most effective classifier and the most important predictors of mental health risk, we hope to provide valuable insights that can be used to improve the support and resources available to university students in the future. Ultimately, the goal of this research is to improve the well-being of university students by identifying and addressing potential mental health concerns early on.

4. Data Collection

The dataset used in this assignment is acquired from Kaggle titled "Student Mental health". The dataset contains data of 101 university students from a local university. The dataset contains related information from the students with a total of 10 variables. The dataset was collected by a survey aimed to gather information about the students' current academic situation and mental health.

<u>Contents of the dataset:</u>

1. "Student Mental health" dataset consists of 10 variables which includes 101 observations which includes:
    a. Timestamp
    b. Choose your gender
    c. Age
    d. What is your course?
    e. Your current year of Study
    f. What is your CGPA?
    g. Marital status
    h. Do you have Depression?
    i. Do you have Anxiety?
    j. Do you have Panic attack?
    k. Did you seek any specialist for a treatment?
2. The data of students' mental health were collected in the midst of covid during 2020.



	Timestamp	Choose your gender	Age	What is your course?	Your current year of Study	What is your CGPA?	Marital status	Do you have Depression?	Do you have Anxiety?	Do you have Panic attack?	Did you seek any specialist for a treatment?
0	8/7/2020 12:02	Female	18.0	Engineering	year 1	3.00 - 3.49	No	Yes	No	Yes	No
1	8/7/2020 12:04	Male	21.0	Islamic education	year 2	3.00 - 3.49	No	No	Yes	No	No
2	8/7/2020 12:05	Male	19.0	BIT	Year 1	3.00 - 3.49	No	Yes	Yes	Yes	No
3	8/7/2020 12:06	Female	22.0	Laws	year 3	3.00 - 3.49	Yes	Yes	No	No	No
4	8/7/2020 12:13	Male	23.0	Mathemathics	year 4	3.00 - 3.49	No	No	No	No	No
...	...	...	...	...	...	...	...	...	...	...	...
96	13/07/2020 19:56:49	Female	21.0	BCS	year 1	3.50 - 4.00	No	No	Yes	No	No
97	13/07/2020 21:21:42	Male	18.0	Engineering	Year 2	3.00 - 3.49	No	Yes	Yes	No	No
98	13/07/2020 21:22:56	Female	19.0	Nursing	Year 3	3.50 - 4.00	Yes	Yes	No	Yes	No
99	13/07/2020 21:23:57	Female	23.0	Pendidikan Islam	year 4	3.50 - 4.00	No	No	No	No	No
100	18/07/2020 20:16:21	Male	20.0	Biomedical science	Year 2	3.00 - 3.49	No	No	No	No	No

101 rows × 11 columns

Missing value and diversification of data format in students' mental health dataset.

As we can observe in the figure above, the dataset requires data preprocessing in order to proceed the data finding as the dataset contains plenty of diversification of data format in some of the columns.

5. Data Pre-processing (Data Cleaning)

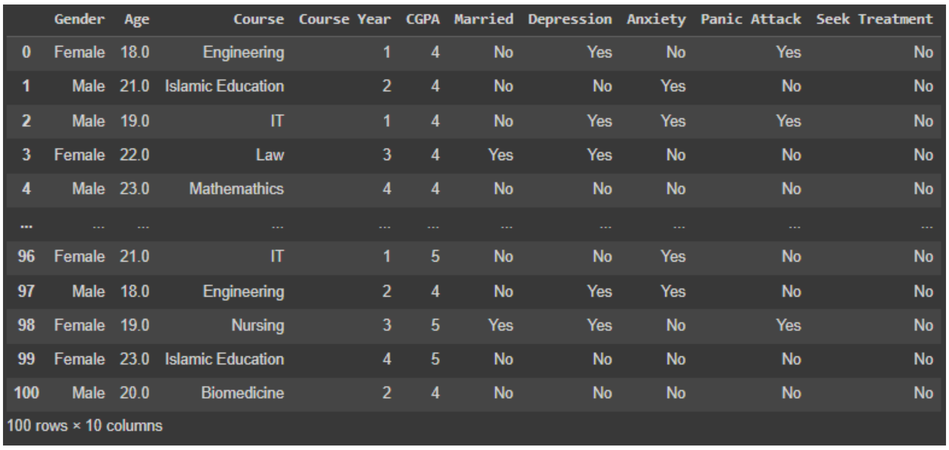
The initial step of this process is to detect rows containing null values in the dataset. After locating the null value, it was discovered that just the 'Age' column in row 43 was affected. Since there was no method to validate the student's age, I decided to delete this row.

The next stage in our procedure would be to remove the column titled "Timestamp," as it will be unnecessary for this experiment. Then, I shortened the names of each column's properties to make them easier to comprehend and to refer to when performing data analysis. The columns include ‘Gender’, ‘Course’, ‘Course Year’, ‘CGPA’, ‘Married’, ‘Depression’, ‘Anxiety’, ‘Panic Attack’ and ‘Seek Treatment’.

Additionally, values in the "Course Year" column were found not consistently formatted, with some entries being written as "year 1" and others as "Year 1". To standardise the format, a custom function was created to extract the last character of each string, which corresponds to the year number. The function also converted the values into integers. This cleaning process ensured that the "Course Year" column was consistent and in a format suitable for analysis. Furthermore, it also allowed for accurate calculations and comparisons to be made based on the course year of the students.

In addition, cleaning the values in the 'Course' column would consume the majority of my work, given the dataset was received from a Google form and student responses are short form. I conducted research on each course in this column to harmonise the data and identify similar courses. An example of this would be ‘engin’ meant Engineering and ‘KOE’ meant Kulliyah of Engineering which will be in the same group as Engineering.

Furthermore, I removed unnecessary space in ‘CGPA’ values and replaced the CGPA scores ranging from 1 to 5, with 1 being the lowest and 5 the highest, in order to facilitate the prediction model and subsequent data analysis.

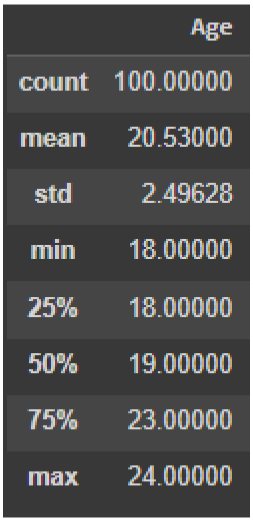


	Gender	Age	Course	Course Year	CGPA	Married	Depression	Anxiety	Panic Attack	Seek Treatment
0	Female	18.0	Engineering	1	4	No	Yes	No	Yes	No
1	Male	21.0	Islamic Education	2	4	No	No	Yes	No	No
2	Male	19.0	IT	1	4	No	Yes	Yes	Yes	No
3	Female	22.0	Law	3	4	Yes	Yes	No	No	No
4	Male	23.0	Mathemathics	4	4	No	No	No	No	No
...	...	...	...	...	...	...	...	...	...	...
96	Female	21.0	IT	1	5	No	No	Yes	No	No
97	Male	18.0	Engineering	2	4	No	Yes	Yes	No	No
98	Female	19.0	Nursing	3	5	Yes	Yes	No	Yes	No
99	Female	23.0	Islamic Education	4	5	No	No	No	No	No
100	Male	20.0	Biomedicine	2	4	No	No	No	No	No

100 rows × 10 columns

Result from data pre-processing

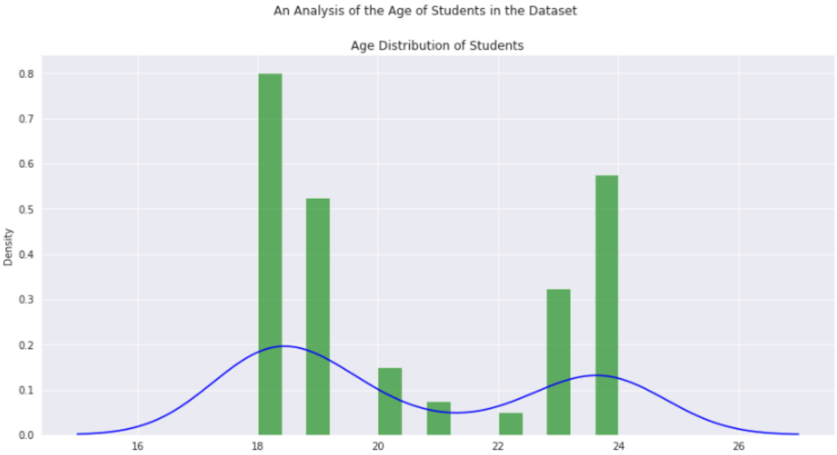
6. Exploratory Data Analysis



	Age
count	100.00000
mean	20.53000
std	2.49628
min	18.00000
25%	18.00000
50%	19.00000
75%	23.00000
max	24.00000

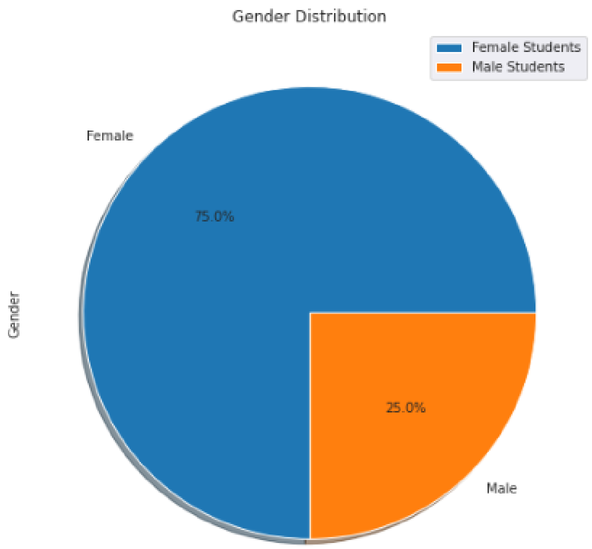
Descriptive Statistics of Age

From the descriptive statistics above, we can clearly observe the count, mean, standard deviation, min, interquartile range and the max value of the student's age. We are able to see that the count is 100 as we have 100 rows of data after performing data preprocessing. The minimum age from our dataset is 18 years old which suggests the youngest of respondents to the survey was a first year student while the max is 24 years old which I would assume is a year 4 student.



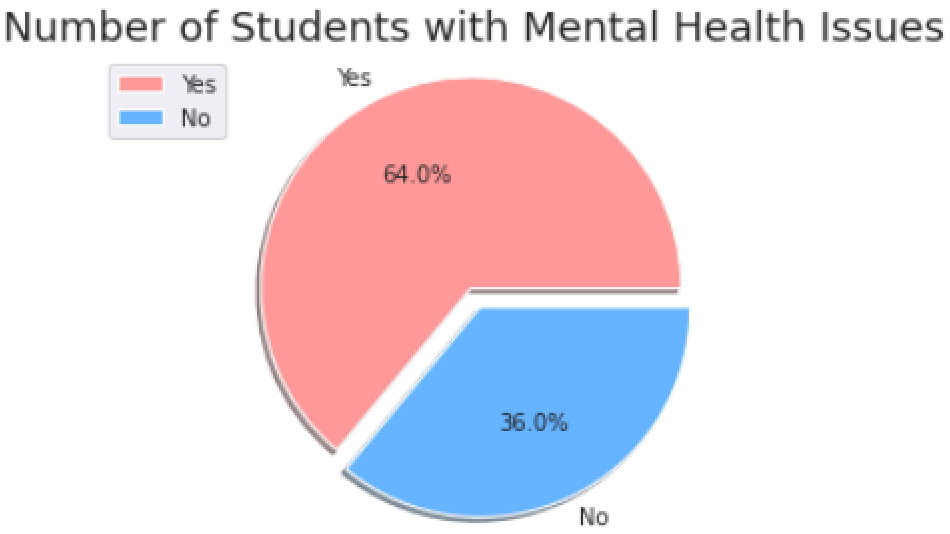
Age distributions of students in the dataset

By looking at the bar chart, we can tell that the respondents are students between the ages of 18 and 24. In comparison to others, responses from students aged 20-23 are minimal.



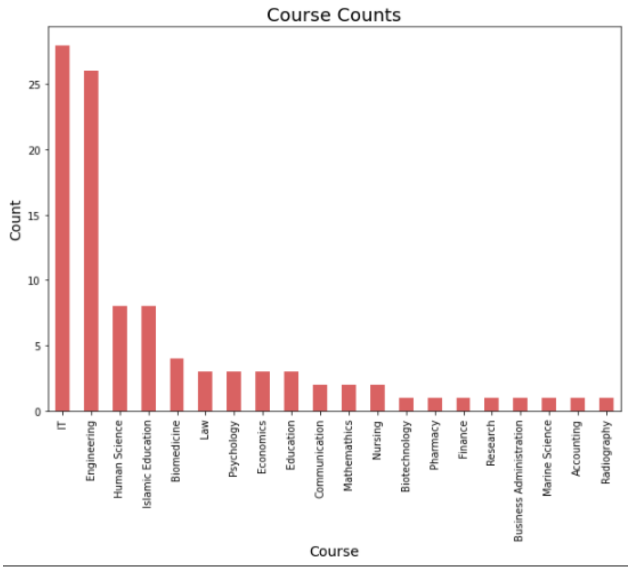
“What is the gender distribution of students in the dataset?”

Majority of the students who took the survey were females which stands at 75% of the dataset while only 25% of male students took the survey.



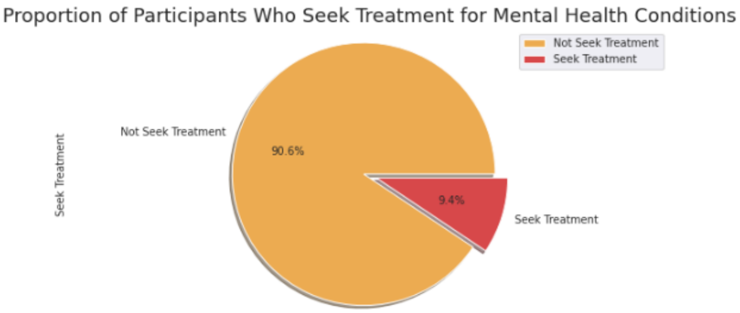
“What is the distribution of students with positive mental health issues and negative mental issues?”

By observing the pie chart above, we can see that 64% of students from the dataset is facing mental health issues and 36% of students does not face either depression, anxiety or panic attack.



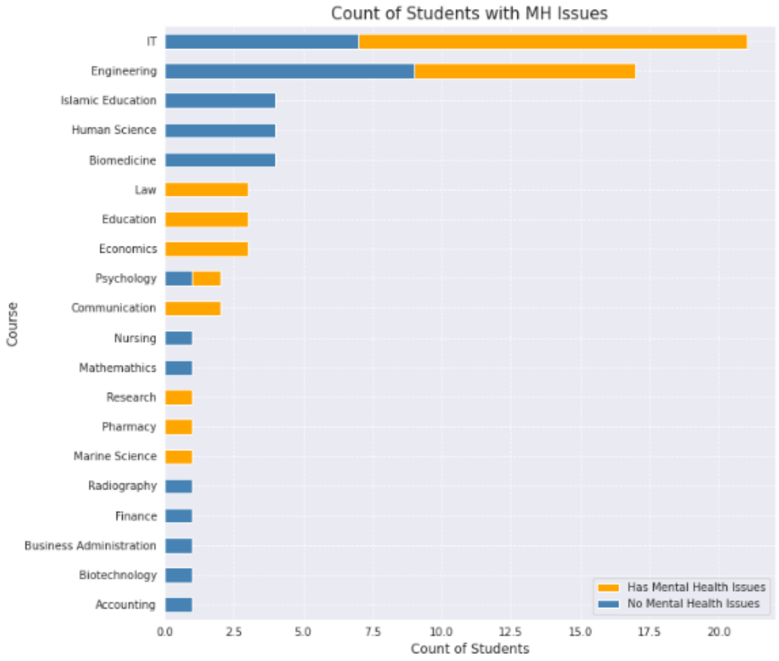
“What is the most common course of the participants in the dataset?”

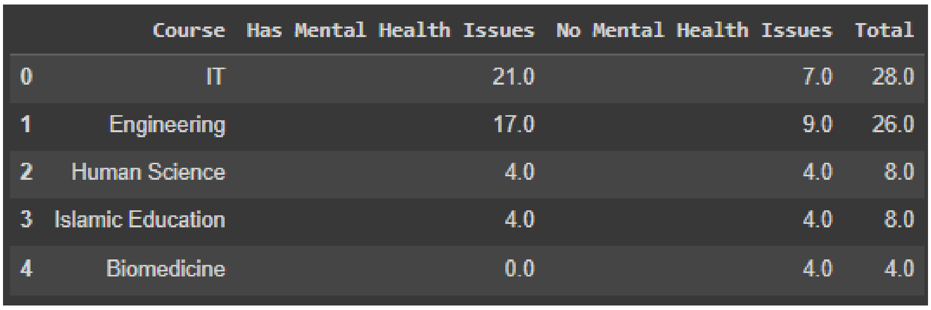
In the bar plot of the dataset, it was observed that IT is the most common course among the student respondents, with the second highest being engineering. Biotechnology, pharmacy, finance, research, business administration, marine science, accounting and radiography were the least common courses among the student respondents. This information provides valuable insight into the composition of the student population in the dataset and may be useful in further analysis and understanding the study population.



"What is the proportion of participants who seek treatment while having mental health conditions in the dataset?"

Based on the pie chart, it can be observed that a vast majority (90.6%) of the student respondents in the dataset do not seek treatment for their mental health conditions. This indicates that a significant portion of students may not be receiving the support they need to manage their mental health. The proportion of students who do seek treatment for their mental health issues (9.4%) is relatively small, highlighting the need for more effective outreach and support for students struggling with mental health issues.

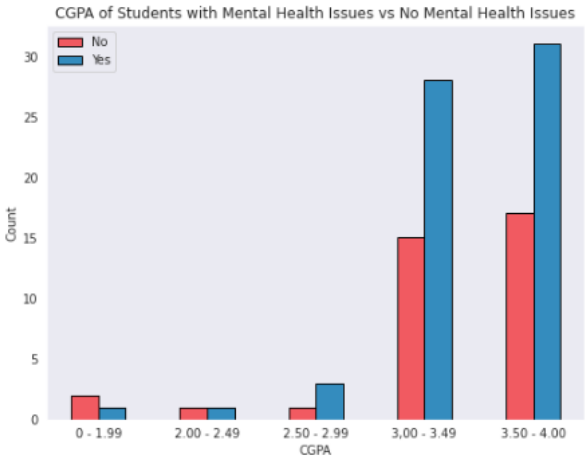




	Course	Has Mental Health Issues	No Mental Health Issues	Total
0	IT	21.0	7.0	28.0
1	Engineering	17.0	9.0	26.0
2	Human Science	4.0	4.0	8.0
3	Islamic Education	4.0	4.0	8.0
4	Biomedicine	0.0	4.0	4.0

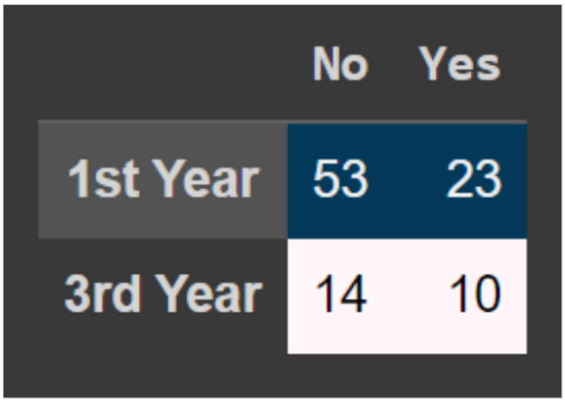
"What is the relationship between the course and the prevalence of mental health issues among students?"

The bar plot of count of students with mental health issues reveals some interesting insights. IT, Engineering, Human Science, Islamic Education, are the top four courses where students report experiencing mental health issues. IT has the highest number of students reporting mental health concerns with 21 out of 27 students. Engineering comes in second with 17 out of 26 students. Human Science and Islamic Education have a similar number of students with 4 out of 8 students each. It is worth noting that these courses tend to be demanding, which could explain why students in these fields are more likely to report mental health issues. This highlights the importance of providing support and resources for students in these fields.

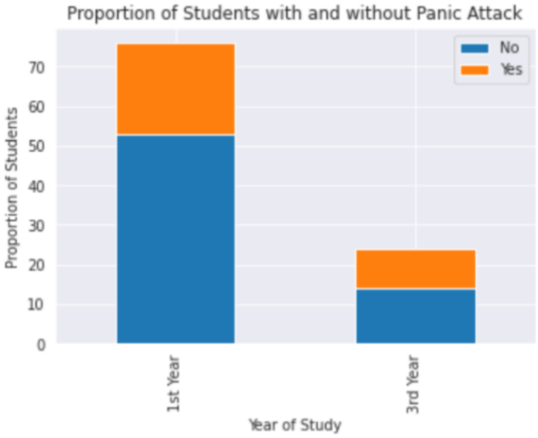


"Is there a significant difference in the CGPA of students with mental health issues compared to students without mental health issues?"

Upon examination of the comparative bar plot of CGPA for students with and without mental health issues, we find that the majority of students have reported a CGPA of 3.0 and above. However, what is intriguing is that despite experiencing mental health issues, the distribution of CGPA for these students is not vastly different from that of their peers who do not report such issues. This suggests that, on average, students with mental health issues are still performing academically at a similar level to their peers.

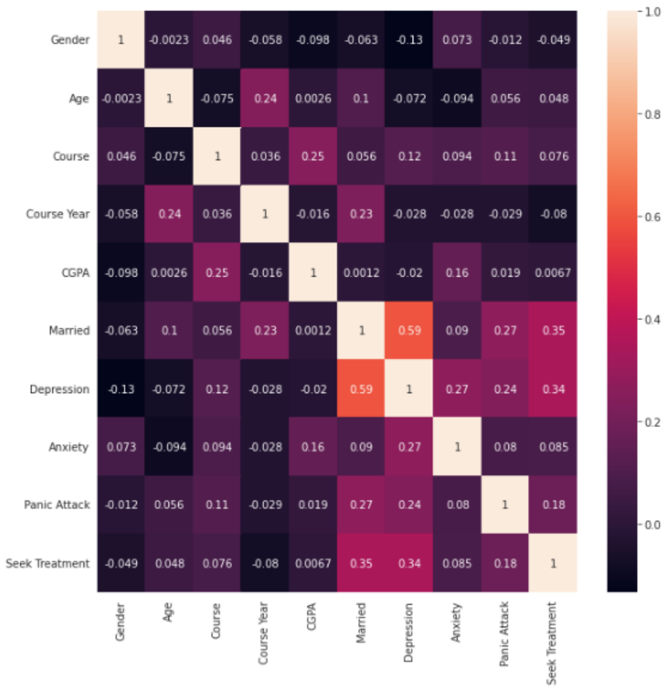


	No	Yes
1st Year	53	23
3rd Year	14	10



“Will a student in their 3rd year of study have a higher likelihood of experiencing a panic attack compared to a student in their 1st year of study?”

Based on the data analysed, it appears that there is not a significant difference in the likelihood of experiencing a panic attack between students in their 1st year of study and students in their 3rd year of study. However, it is worth noting that while a higher proportion of 1st year students (around 70%) did not report experiencing a panic attack, the sample size of 3rd year students is smaller, which means 3rd year students have a higher likelihood of experiencing a panic attack. This may be due to 3rd year students being close to graduating and stressing on harder subjects.



Heatmap on correlation analysis of the dataset

The heatmap of correlation analysis between variables in our dataset reveals some intriguing insights. Perhaps the most striking discovery is the strong association between Marital Status and Depression. This suggests that marital status may play a significant role in the development of depression symptoms among university students. Furthermore, the heatmap also highlights a strong correlation between Anxiety, Panic Attack, and Depression. This highlights the interconnected nature of these mental health issues and the importance of addressing them together. Additionally, the heatmap also indicates a slight correlation between Marital Status and seeking medical assistance (treatment) which implies that marital status may also be a factor in the decision to seek help for mental health issues among university students. Overall, these findings provide valuable insights into the underlying factors that contribute to mental health issues among university students.

7. Feature Selection

"What is the suitable feature selection technique to use for predicting the cause of mental health?"

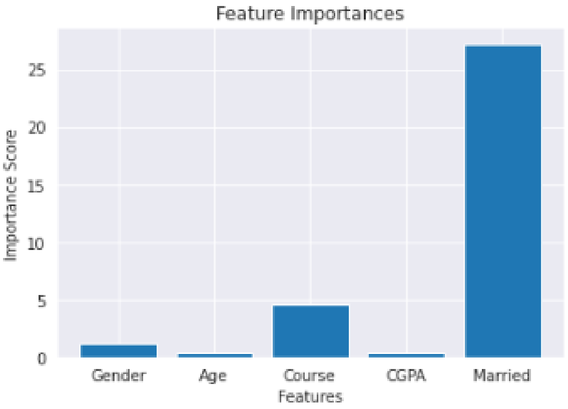
The suitable feature selection technique to use for predicting the cause of mental health is the chi-squared test in combination with the SelectKBest function and k-means clustering. The chi-squared test is a statistical test that calculates the relationship between each feature and the target variable, and scores them based on how informative they are in predicting the target variable. This allows us to determine which factors are most relevant in predicting mental health outcomes.

Additionally, the SelectKBest function allows us to select the top k features with the highest scores from the chi-squared test, which helps in reducing the dimensionality of the data, and improves the accuracy and efficiency of the final classifier.

Furthermore, the k-means clustering algorithm is particularly useful in identifying patterns and structures in the data, which can be helpful in understanding the causes of mental health issues.

"How should I obtain the optimal feature set?"

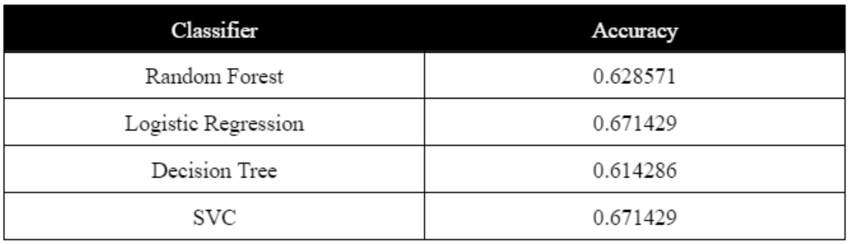
The optimal feature set can be obtained by using feature selection techniques. In this topic's case, chi-squared test and the feature selection algorithm used is SelectKBest. The feature selection selects the top 5 features that have highest correlation with the response variable. The descriptor variable would be the student's gender, age, course, course year, CGPA and marital status. The response variable is Depression, Anxiety, Panic Attack and Mental Health Issue.



Bar plot for feature importances

Conclusion from feature selection: Gender, Age, Course, CGPA and Marital Status are believed to have the most impact on mental health issues columns in the dataset.

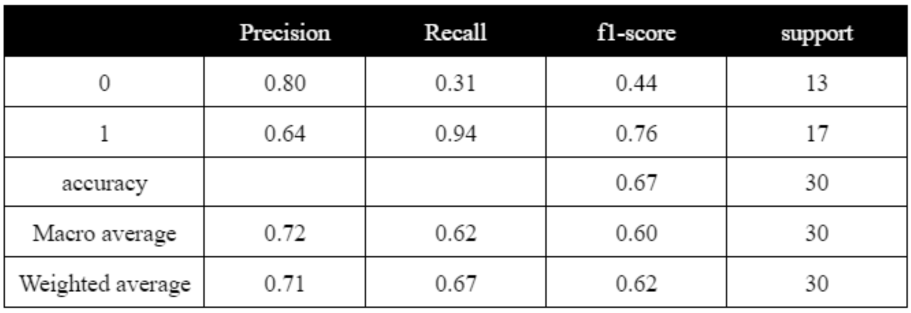
8. Model Selection and Comparison



Classifier	Accuracy
Random Forest	0.628571
Logistic Regression	0.671429
Decision Tree	0.614286
SVC	0.671429

Accuracy score for four classifiers

Logistic Regression Classifier appears to be the most accurate among 4 classifiers in our case where the features used are Gender, Age, Course, CGPA and Marital Status selected from feature selection, which are all factors that have been shown to be related to mental health issues.



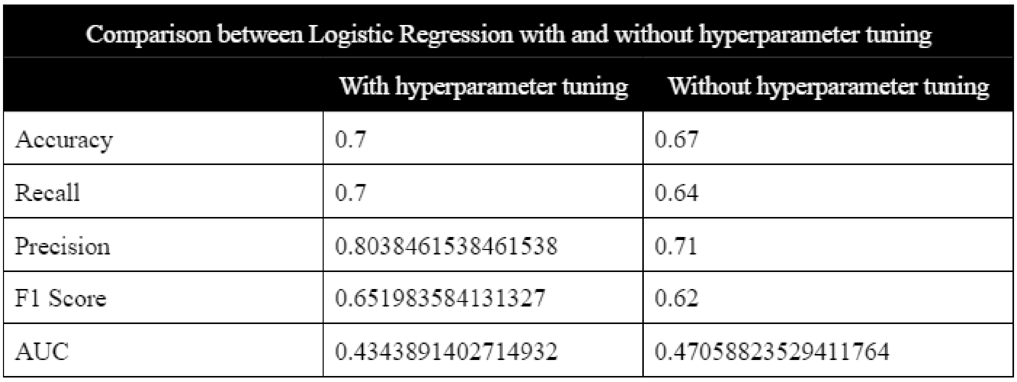
	Precision	Recall	f1-score	support
0	0.80	0.31	0.44	13
1	0.64	0.94	0.76	17
accuracy			0.67	30
Macro average	0.72	0.62	0.60	30
Weighted average	0.71	0.67	0.62	30

Classification report for Linear Regression

The classification report for Logistic Regression Classifier shows precision, recall, f1-score, and support for each class. For class 0 (no mental health issues), 80% of predictions were correct and 31% of actual cases were identified. The f1-score for class 0 is 0.44. For class 1 (mental health issues), 64% of predictions were correct and 94% of actual cases were identified. The f1-score for class 1 is 0.76. Overall, the model has a good performance in identifying mental health issues, with a recall of 0.94 and f1-score of 0.76.

In conclusion, the performance of this linear regression model is an acceptable model for identifying mental health issues as it is performing relatively well in identifying cases of mental health issues, with a recall of 0.94 and an f1-score of 0.76.

Hyperparameter tuning



Comparison between Logistic Regression with and without hyperparameter tuning		
	With hyperparameter tuning	Without hyperparameter tuning
Accuracy	0.7	0.67
Recall	0.7	0.64
Precision	0.8038461538461538	0.71
F1 Score	0.651983584131327	0.62
AUC	0.4343891402714932	0.47058823529411764

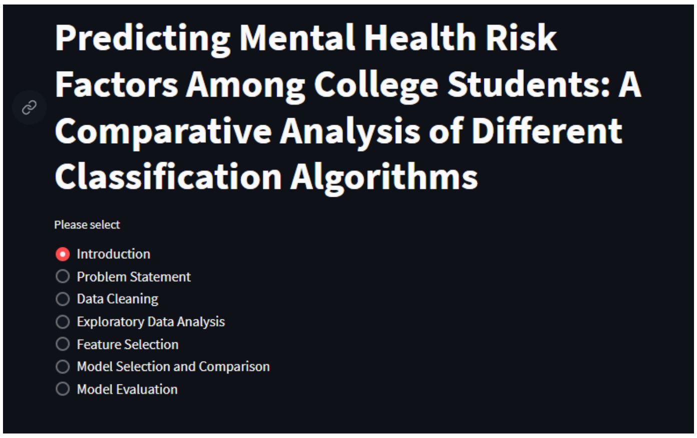
Comparison between Linear Regression with and without hyperparameter tuning

Hyperparameter tuning was performed using GridSearchCV from sklearn.model_selection library. The range of values for the 'C' and 'penalty' hyperparameters were [0.1, 1, 10, 100, 1000] and ['l1', 'l2'] respectively. The evaluation metric used to select the best set of hyperparameters was accuracy. The best set of hyperparameters found were C = 0.1 and penalty = 'l2' by using a 5-fold cross validation technique.

The model with hyperparameter tuning had higher accuracy, recall, precision, and F1 scores, but a lower AUC score than the model without tuning. It's crucial to balance recall and AUC when identifying students with mental health issues. Overall, the model with tuning performed better, but the trade-off between AUC and other metrics should be considered.

9. Deployment

For deployment, I am using Streamlit. The streamlit app is being hosted by using Streamlit cloud. A github account is needed to upload my files by creating a repository. Besides a github account, a streamlit cloud account is needed as well.



Predicting Mental Health Risk Factors Among College Students: A Comparative Analysis of Different Classification Algorithms

Please select

● Introduction
○ Problem Statement
○ Data Cleaning
○ Exploratory Data Analysis
○ Feature Selection
○ Model Selection and Comparison
○ Model Evaluation

My streamlit web app can be accessed using this link:
https://danielteng520-datasciencefundamentals-1191201519-1g7d7g.streamlit.app/

10. Limitations and future proposals

In this assignment, I have applied logistic regression to predict the cause of mental health issues among university students. I have used 70% of the data selected by feature selection for training and the remaining 30% for testing.

In summary, the logistic regression model with hyperparameter tuning has a slightly better performance in predicting the cause of mental health issues among university students, with an accuracy of 70% compared to 67% without tuning.

In conclusion, as this study was conducted on a small dataset of only 100 university students, it is hard to generalise the findings to the entire population. Furthermore, the dataset is also relatively limited in terms of the number of features that were collected, which may have limited the ability to identify the underlying causes of mental health issues.

Future Work

To improve the understanding of the factors that contribute to mental health among university students, a larger and more comprehensive dataset would be needed. This could include a wider range of demographic, academic, and mental health-related variables, as well as more detailed information about the specific causes and symptoms of mental health issues. Additionally, other machine learning techniques such as Decision Tree, Support Vector Machine etc should be applied to the dataset to compare and contrast the results obtained from the logistic regression model.