



A major project report on

## **MENTAL HEALTH PREDICTION IN STUDENTS**

submitted in partial fulfillment of the requirements for the degree of

B. Tech

In

Electronics and Telecommunication Engineering

By

<b>ANIS GHOSH</b>	<b>1804430</b>
<b>HARSH BANSAL</b>	<b>1804442</b>
<b>SUMIT SAHA</b>	<b>1804485</b>
<b>TINA SASMAL</b>	<b>1804486</b>
<b>UTSAB GHOSH</b>	<b>1804488</b>
<b>SIRAK GUHA NIYOGI</b>	<b>1804477</b>

under the guidance of

**Prof. V K Shrivastava**

School of Electronics Engineering  
**KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**  
(Deemed to be University)  
BHUBANESWAR

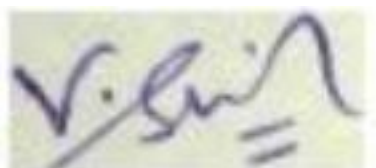
APRIL 2022

## CERTIFICATE

This is to certify that the project report entitled “**MENTAL HEALTH PREDICTION IN STUDENTS**” submitted by

<b>ANIS GHOSH</b>	<b>1804430</b>
<b>HARSH BANSAL</b>	<b>1804442</b>
<b>SUMIT SAHA</b>	<b>1804485</b>
<b>TINA SASMAL</b>	<b>1804486</b>
<b>UTSAB GHOSH</b>	<b>1804488</b>
<b>SIRAK GUHA NIYOGI</b>	<b>1804477</b>

in partial fulfilment of the requirements for the award of the **Degree of Bachelor of Technology in Electronics and Telecommunication Engineering** is a bonafide record of the work carried out under my(our) guidance and supervision at School of Electronics Engineering, KIIT (Deemed to be University).



Signature of Supervisor 1  
Prof. V K Shrivastava  
School of Electronics Engineering  
KIIT (Deemed to be University)

**The Project was evaluated by us on \_\_\_\_\_**

EXAMINER 1  
EXAMINER 3

EXAMINER 2  
EXAMINER 4

## ACKNOWLEDGEMENTS

We are overjoyed and honoured to be able to express our heartfelt gratitude to our supervisor, Professor V K Shrivastava, for his excellent guidance throughout our project work. We have been inspired by his kindness, dedication, hard work, and meticulous attention to detail. Thank you so much, sir, for your unwavering support and patience with us. We'd like to express our gratitude to him for patiently and meticulously correcting all of our manuscripts. We appreciate the help we received from the Web Scrapping Organization website, which provided us with a wealth of information about the modules.

### STUDENT SIGNATURE

Roll Number	Name	Signature
1804430	ANSI GHOSH	
1804442	HARSH BANSAL	
1804485	SUMIT SAHA	
1804486	TINA SASMAL	
1804488	UTSAB GHOSH	
1804477	SIRAK GUHA NIYOGI	

Date:- 01.04.22

## **ABSTRACT**

The language of deficits and problems, rather than resources and strengths, is used in behavioural health treatment, which is based on the medical model. It is now possible to refocus on well-being rather than illness, thanks to advances in the field of positive psychology. Mental health issues, as well as personality disorders, have been shown in studies to have a negative impact on academic performance. Positive mental health, on the other hand, can act as a protective shield against mental illness. The Mental Health data set, which is available in the source repository, is used in this paper.

# TABLE OF CONTENTS

Abstract

Table of Contents

List of Figures

List of Tables

List of symbols/ abbreviations

CHAPTER I: INTRODUCTION 12

1.1 Overview 12

CHAPTER 2: DATASET METHODOLOGY 14

2.1 Dataset Exploration 14

2.2 Technology Used 14

2.3 Proposed Working 15

2.4 The working Model 16

2.5 Limitations 16

CHAPTER 3: BACKGROUND/BASIC CONCEPTS 17

3.1 Data Analysis 17

3.2 Classification Algorithms used 20

3.2.1. Logistic Regression

3.2.2. KNN

3.2.3. SVM

3.2.4. Naive Bayes Classifier

3.2.5. Decision Tree Classifier

3.3 Result Algorithms Used 26

3.4	Feature Selection	26
	3.4.1. Extra Tree Classifier	
	3.4.2. Heat Map	
	3.4.3. Chi-Square Test	
3.5	Analyzing the Output	31
	3.5.1 Confusion Matrix	
3.6	Observations	34

## CHAPTER 4: Literature Review

4.1	A systematic review of studies of depression	40
	4.1.1 Methodology	
	4.1.2 Conclusion	
4.2	Development of an accumulative stress scale	42
	4.2.1 Methodology	
	4.2.2 Conclusion	
4.3	Depression among Chinese international students	44
	4.3.1 Methodology	
	4.3.2 Conclusion	
4.4	Depression among Japanese international students	46
	4.4.1 Methodology	
	4.4.2 Conclusion	
4.5	Psychological and socio-economic factors	49
	4.5.1 Methodology	
	4.5.2 Conclusion	

REFERENCES	53
------------	----

# LIST OF FIGURES

<b>Fig no.</b>	<b>Description</b>	<b>Page No.</b>
2.1	Dataset Preview	14
2.2	Flowchart representation of Proposed Working model.	15
2.3	Design of the FEPSY Study	16
3.1	Linear Classifiers and their Usage	21
3.2	A Journey from Decision Function to Decision Boundary	21
3.3	Image showing how similar data points typically exist close to each other	22
3.4	SVM Generalized	23
3.5	SVM Mechanism	23
3.6	Bayes Theorem Equation	24
3.7	Decision Tree Classifier generalized	25
3.8	Visualization of a Random Forest Model Making a Prediction	26
3.9	Feature Selection for CASE A	28
3.10	Feature Selection for CASE B	28

<b>Fig no.</b>	<b>Description</b>	<b>Page No.</b>
3.8	Visualization of a Random Forest Model Making a Prediction	26
3.9	Feature Selection for CASE A	28
3.10	Feature Selection for CASE B	28
3.11	Feature Selection for CASE C	29
3.12	Heat Map	30
3.13	Formula for chi square	31
4.1	Methodology flowchart for Journal of psychiatric research, 47(3), 391-400.	39
4.2	Accumulative stress scale, SCS, PHQ-9 evaluation formulation	45
4.3	Descriptive statistics for contentious variable	48



# LIST OF TABLES

<b>Fig no.</b>	<b>Description</b>	<b>Page No.</b>
3.1	Logistic Regression accuracy results for Case A	23
3.2	KNN accuracy results for Case A	23
3.3	Decision Tree Classifier accuracy results for Case A	23
3.4	SVM accuracy results for Case A	24
3.5	Naive Bayes accuracy results for Case A	24
3.6	Random Forest Classifier accuracy results for Case A	24
3.7	Logistic Regression accuracy results for Case B	25
3.8	KNN accuracy results for Case B	25
3.9	Decision Tree Classifier accuracy results for Case B	25
3.10	SVM accuracy results for Case B	26
3.11	Naive Bayes accuracy results for Case B	26
3.12	Random Forest Classifier accuracy results for Case B	26
3.13	Logistic Regression accuracy results for Case C	27
3.14	KNN accuracy results for Case C	27
3.15	Decision Tree Classifier accuracy results for Case B	27
3.16	SVM accuracy results for Case C	28
3.17	Naive Bayes accuracy results for Case C	28
3.18	Random Forest Classifier accuracy results for Case C	28

## INTRODUCTION

---

### 1.1. OVERVIEW

Mental health is a state of well-being in which an individual recognises his or her own potential, is able to cope with daily stressors, works successfully and productively, and contributes to his or her community.

Physical and mental health are the result of a complex interaction between a number of individual and environmental factors, including: Family history of illness and disease/genetics Lifestyle and health behaviour (e.g., smoking, exercise, substance use) Levels of personal and workplace stress Exposure to toxins Exposure to trauma Personal life circumstances and history.

### 1.1. The Instigation Process

Mental health suffers when a person's resources and coping abilities are pushed to their limits. Working long hours in inclement weather and caring for a chronically ill relative are two examples of common responsibilities. Unemployment, underemployment, and poverty can all be harmful to one's mental health.

Mental illness is a medically recognised condition in which a person's cognitive, affective, or relational abilities are severely impaired. Mental disorders are caused by biological, developmental, and/or psychosocial factors, and they can be treated similarly to physical illnesses. Medical experts have accumulated a large amount of medical data that can be analysed and valuable information extracted.

Approaches for extracting meaningful and hidden information from huge volumes of data are known as machine learning techniques. In a medical database, the bulk of the information is discrete. As a result, making judgments based on discrete facts is a challenging task.

Data mining branch Machine Learning (ML) excels at managing large, well-formatted datasets. In the medical field, machine learning may be used to diagnose, detect, and forecast a variety of ailments. The main goal of this study is to provide doctors with a tool to aid in the early detection of mental diseases. As a consequence, it will be simpler to provide patients with proper medication while minimising major side effects. Machine learning plays a crucial role in discovering hidden discrete patterns and interpreting data.

Machine learning techniques help in the prediction and early detection of heart illness after data analysis. The efficiency of different machine learning approaches, such as Logistic Regression and Support Vector Machine, in detecting mental illness early on and reducing suicidal impulses in people is investigated in this study.

## DATASET METHODOLOGY

### 2.1. Dataset Exploration

There are 286 rows and 50 columns in this dataset. The information was gathered from a group of domestic and international students at Ritsumeikan Asia Pacific University who completed the General Health Help-Seeking Questionnaire and Patient Health Questionnaire, both of which were based on the Acculturative Stress Scale for International Students and the Social Connectedness Scale. There were a few missing values that were dealt with and handled well. Researchers and anyone working in the medical profession can use this dataset. (Source: <http://www.mdpi.com/2306-5729/4/3/124/s1.>)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
	Inter_dom	Region	Gender	Academic	Age	Age_cat	Stay	Stay_Cate	Japanese	Japanese_e	English	English_ca	Intimate	Religion	Suicide	Dep	DepType	ToDep	DepSev	ToSC	APD	AtHome	APH	Afear	ACS	AGuilt	AMiscell	ToAS	Partner	Friends
2	Inter	SEA	Male	Grad	24	4	5	Long	3	Average	5	High	Yes	No	No	No	No	0	Min	34	23	9	11	8	11	2	27	91	5	
3	Inter	SEA	Male	Grad	28	5	1	Short	4	High	4	High	Yes	No	No	No	No	2	Min	48	8	7	5	4	3	2	10	39	7	
4	Inter	SEA	Male	Grad	25	4	6	Long	4	High	4	High	Yes	Yes	No	No	No	2	Min	41	13	4	7	6	4	3	14	51	3	
5	Inter	EA	Female	Grad	29	5	1	Short	2	Low	3	Average	No	No	No	No	No	3	Min	37	16	10	10	8	6	4	21	75	5	
6	Inter	EA	Female	Grad	28	5	1	Short	1	Low	3	Average	Yes	No	No	No	No	3	Min	37	15	12	5	8	7	4	31	82	5	
7	Inter	SEA	Male	Grad	24	4	6	Long	3	Average	4	High	Yes	No	No	No	No	6	Mild	38	18	8	10	8	7	3	29	83	6	
8	Inter	SA	Male	Grad	23	4	1	Short	3	Average	5	High	Yes	No	No	No	No	3	Min	46	17	6	10	5	3	2	15	58	7	
9	Inter	SEA	Female	Grad	30	5	2	Medium	1	Low	1	Low	Yes	Yes	Yes	No	No	9	Mild	41	16	20	19	15	11	6	40	127	7	
10	Inter	SEA	Female	Grad	25	4	4	Long	4	High	4	High	No	No	No	Yes	Other	7	Mild	36	22	12	13	13	10	6	33	109	4	
11	Inter	Others	Male	Grad	31	5	2	Medium	1	Low	4	High	Yes	Yes	No	No	No	4	High	48	8	4	5	12	3	2	17	51	1	
12	Inter	Others	Female	Grad	28	5	1	Short	1	Low	2	Low	No	Yes	No	No	No	5	Mild	32	24	8	10	8	6	6	30	92	7	
13	Inter	SEA	Female	Grad	31	5	1	Short	1	Low	4	High	Yes	Yes	No	No	No	8	Mild	47	17	12	14	13	9	4	26	95	6	
14	Inter	SA	Male	Grad	29	5	1	Short	1	Low	4	High	Yes	Yes	No	No	No	1	Min	48	8	11	5	4	7	2	17	54	7	
15	Inter	EA	Male	Grad	23	4	1	Short	3	Average	4	High	Yes	Yes	No	No	No	3	Min	32	9	8	5	4	6	7	18	57	7	
16	Inter	SEA	Female	Grad	31	5	1	Short	1	Low	3	Average	Yes	No	Yes	No	Yes	9	Mild	31	23	16	15	8	12	8	30	112	1	
17	Inter	Others	Female	Grad	30	5	1	Short	1	Low	5	High	Yes	Yes	No	Yes	Other	6	Mild	40	19	9	5	4	13	2	22	74	7	
18	Inter	EA	Female	Grad	31	5	1	Short	1	Low	3	Average	Yes	No	No	No	No	3	Min	48	11	13	7	4	8	5	15	63	6	
19	Inter	Others	Female	Grad	29	5	1	Short	1	Low	5	High	Yes	Yes	No	No	No	3	Min	48	16	4	5	4	3	2	13	47	7	
20	Inter	SEA	Female	Under	19	2	1	Short	3	Average	5	High	No	No	No	No	No	7	Mild	44	11	8	5	7	4	2	18	55	5	
21	Inter	SEA	Male	Under	25	4	3	Medium	2	Low	4	High	Yes	No	No	No	No	1	Min	36	13	10	7	5	3	3	19	60	1	
22	Inter	SEA	Male	Under	18	1	1	Short	1	Low	4	High	No	No	No	No	No	4	Min	26	18	4	7	4	4	2	27	66	6	
23	Inter	SEA	Male	Under	18	1	1	Short	2	Low	3	Average	No	No	No	No	No	3	Min	26	11	5	5	7	3	2	33	66	2	
24	Inter	SEA	Male	Under	19	2	1	Short	1	Low	4	High	Yes	Yes	Yes	Yes	Other	13	Mod	25	25	16	10	11	6	4	22	94	1	
25	Inter	SEA	Male	Under	20	2	3	Medium	3	Average	4	High	No	No	No	No	No	1	Min	34	24	16	15	12	11	8	37	123	6	
26	Inter	EA	Female	Under	29	5	1	Short	1	Low	3	Average	Yes	No	No	Yes	Other	8	Mild	39	18	8	10	8	7	4	23	78	2	
27	Inter	Others	Female	Grad	25	4	8	Long	4	High	4	High	No	Yes	Yes	Other	10	Mod	44	12	9	9	8	5	3	19	65	2		
28	Inter	SEA	Female	Grad	30	5	2	Medium	3	Average	4	High	Yes	Yes	No	Yes	Other	13	Mod	42	23	15	11	9	5	2	26	91	7	
29	Inter	SEA	Female	Under	17	1	1	Short	2	Low	3	Average	Yes	Yes	Yes	No	No	9	Mild	38	19	9	10	11	7	6	26	88	2	
30	Inter	SEA	Female	Under	20	2	2	Medium	3	Average	4	High	No	Yes	No	No	No	6	Mild	46	10	4	5	4	3	2	11	39	1	
31	Inter	EA	Male	Under	19	2	1	Short	3	Average	3	Average	No	No	No	No	No	7	Mild	40	39	4	5	4	3	2	15	72	2	
32	Inter	EA	Female	Under	18	1	1	Short	3	Average	3	Average	No	No	No	Yes	Other	10	Mod	38	16	9	13	6	3	2	14	66	2	
33	Inter	SA	Female	Under	21	3	2	Medium	3	Average	4	High	Yes	Yes	Yes	No	No	9	Mild	43	16	8	10	8	6	4	21	73	2	
34	Inter	SA	Male	Under	22	3	2	Medium	3	Average	5	High	No	Yes	Yes	No	No	9	Mild	40	16	9	10	8	6	4	20	73	2	
35	Inter	SEA	Male	Under	20	2	2	Medium	3	Average	4	High	Yes	Yes	No	No	No	2	Min	48	8	7	5	4	5	4	12	45	4	
36	Inter	EA	Male	Under	19	2	1	Short	3	Average	3	Average	No	Yes	Yes	No	No	9	Mild	41	18	8	5	4	3	2	10	50	3	
37	Inter	SEA	Female	Under	21	3	4	Long	2	Low	4	High	No	No	Yes	Yes	Major	14	Mod	35	20	11	5	4	5	2	16	63	5	
38	Inter	SEA	Female	Under	22	3	3	Medium	4	High	4	High	No	Yes	No	No	No	4	Min	44	8	14	5	8	3	4	21	63	5	
39	Inter	EA	Female	Under	18	1	1	Short	3	Average	3	Average	No	No	Yes	Yes	Major	14	Mod	31	14	7	12	9	4	3	19	68	2	
40	Inter	JAP	Female	Under	21	3	10	Long	5	High	3	Average	No	No	Yes	Yes	Other	13	Mod	32	9	4	10	8	4	2	13	50	1	
41	Inter	EA	Female	Under	22	3	3	Medium	3	Average	4	High	Yes	No	No	Yes	Major	15	ModSev	25	23	6	14	12	3	8	26	92	5	
42	Inter	SEA	Male	Under	19	2	1	Short	3	Average	5	High	No	No	No	No	No	6	Mild	44	12	9	5	5	5	3	20	59	3	
43	Inter	EA	Male	Under	24	4	2	Medium	4	High	4	High	No	No	No	No	No	4	Min	46	15	9	15	10	6	4	19	78	3	

Fig 2.1 : Dataset Preview

### 2.2. Technology Used:

Our experiment is based on Supervised Machine Learning, in which both the input and output data (that is, the goal variable) are provided. We obtained our data from a reliable source and then cleaned it by removing any unnecessary elements. In our work, we employed feature selection techniques and then scaled it to fit our needs. We then separated the training and testing data and used the training data to build a model, which we then applied to the testing data. Our efforts generated unprecedented outcomes.

For our experiment, we used Python 3.8 and the Jupyter Notebook IDE.

### 2.3. Proposed Working:

By examining classification algorithms and performing performance analysis, the proposed work predicts whether a person would try suicide based on multiple mental disease restrictions. The goal of this research is to determine whether or not the patient is suffering from mental depression or other diseases. The data is fed into a model that determines whether the person is likely to attempt suicide or not after the health professional measures numerous parameters from the patient's health report. Our experiment is set to follow such a flowchart:

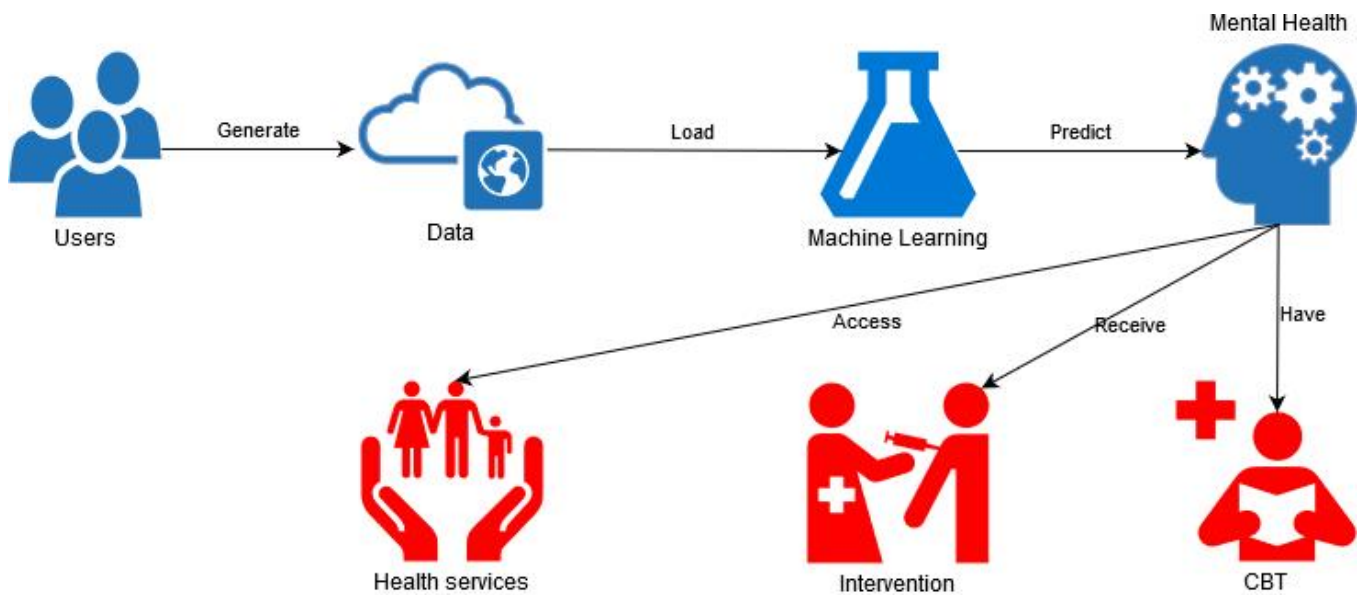


Fig 2.2 : Flowchart representation of Proposed Working model.

## 2.4. The working Model :

Our proposed model can be used in a Basel Psychosis Screening Instrument (BSIP). It's a technology that detects the most significant risk factors for mental illness or suicide ideation. As a result, our model can be used to inspire the architecture of this gadget, which will aid in the detection of mental illness in kids.

The Basel Screening Instrument für Psychosen (BSIP; Riecher-Rössler et al., 2008) [Basel Screening Instrument for Psychosis] [Basel Screening Instrument for Psychosis] [Basel Screening Instrument for Psychosis] was created as a screening tool for detecting individuals with beginning psychosis in the atypical early stages of the disease. The tool was created after a thorough review of the literature on the most relevant risk factors and early symptoms of schizophrenia psychoses. The BSIP is a checklist that includes guidelines for conducting a semi-structured clinical interview and covers seven criteria of importance for detecting psychosis early on: Age; Psychopathology; Kink over time (Loss of social roles); Drug addiction; Childhood mental illnesses/psychological disorders; Genetic risk; Allocation with suspected psychosis To assess interrater reliability, data from 24 psychiatric cases were obtained.

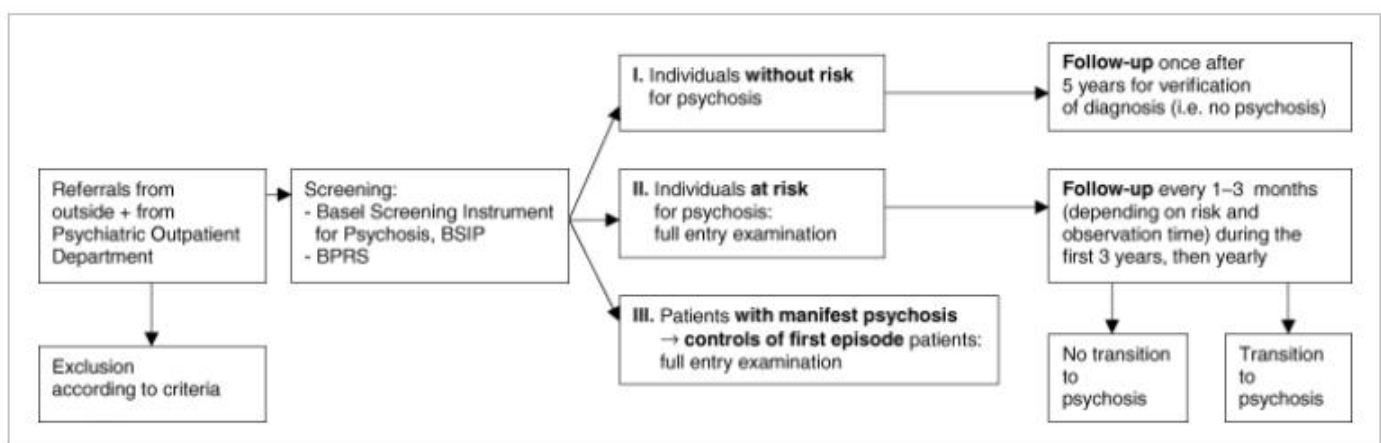


Fig 2.3 :Design of the FEPSY Study.

## 2.5. Limitations :

- Our experiment is just theoretical. As a result, actual applications must be implemented in accordance with the working environment.
- To get the desired results, the input values must be rescaled. It's likely that a larger number of features won't be able to be scaled.
- The experiment is based on a hit-or-miss approach employing several algorithms. As a result, the model may not provide improved accuracy in a certain area.

## CHAPTER 3

### BACKGROUND / BASIC CONCEPTS

---

#### 3.1. Data Analysis

In data mining, exploratory data analysis (EDA) is a process of assessing datasets in order to summarise their important aspects, usually utilising visual approaches. EDA is used to see what the data can tell us before we start modelling. Deducing crucial data attributes from a column of numbers or a complete spreadsheet is difficult. The process of extracting insights from raw data can be tedious, dull, and/or overwhelming. Exploratory data analysis techniques have been developed to assist in this instance.

Exploratory data analysis may be classified in two ways.

- i. First and foremost, each approach is either non-graphical or graphical.
- ii. The fact that each approach is either univariate or multivariate is another consideration (usually just bivariate).

In our experiment, which is a Supervised Machine Learning Problem, both the inputs and the objective variable are supplied. When new examples are provided, supervised learning uses data and associated target answers, which might be numeric values or text labels like classes or tags, to anticipate the correct response.

Exploratory Data Analysis (EDA) is a data analysis approach/philosophy that employs a variety of (mostly graphical) techniques to achieve the following goals: maximise insight into a data set; uncover underlying structure; extract important variables; detect outliers and anomalies; test underlying assumptions; develop parsimonious models.

"Any model, employing any algorithm, before training." Data preprocessing is the most important and will be the most important phase. Several checkpoints (steps) are included in the data preprocessing, including:

**Step 1:** Import Libraries: I utilized Pandas for data manipulation and analysis, Numpy for numerical analysis, Matplotlib, and Seaborn for better data visualization and graphical statistics.

**Step 2:** Import the Dataset: The accuracy of our model is determined by the quantity and quality of your data. We used kaggle to get information.

**Step 3:** Data Preparation: This is where we wrangle the data and get it ready for training. We also clean anything that needs it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.). We can also randomize data, which removes the influence of the order in which our data was acquired and/or otherwise prepared. We, too, can use data visualization to aid in the detection of meaningful correlations between variables or class imbalances (bias alert! ), as well as perform additional exploratory analysis and divide data into training and evaluation sets.

**Step 4:** Checking Null Values: In the previous investigation, we discovered that the number of missing values in most cases surpasses 5% of the data, i.e.,  $5\% \text{ of } 286 = 14.3$ , which exceeds the maximum number of missing values in most circumstances. Furthermore, because this is a Multivariate Analysis, we must take into account all of the columns that contribute to the analysis, and we will not be able to eliminate those columns in the future. We can, however, drop those that are less than that.

**Step 5:** Remove the null values indicated above: After deleting the above missing values from the above exploration, we are left with those missing values that may have a greater impact on our data analysis. However, because many of these data are categorical in nature, we must transform them to numerical representations in order to analyze them further. Many columns were converted here, including 'inter dom,' 'Region,' 'Gender,' 'Academic,' and so on.

**Step 6:** Categorical Data Conversion: In the previous investigation, when we convert our categorical data, our columns are bound to get re-shuffled, and the new numerical columns generated from above are inserted one after the other after the previous indexed 50th column. As a result, we expand the default view scope in our IDE to gain the right perspective, which will aid us in our analysis.



**Step 7:** Replace missing values with NaN: After turning our categorical data into numerical data, the section of missing values is replaced with NaN in the above exploration. 'Not a Number' is the abbreviation for 'Not a Number.' NaN can be substituted with 0 or the average values of all the data in the column for analysis purposes. We figured we'd just replace it with 0 and keep going with the model. It's possible that our assumption is incorrect, which might happen if our model's accuracy falls short of expectations. But for the time being, we'll stick with 0 and move on.

**Step 8:** Creating a Training and Test Set from the Dataset: To train our machine learning model, I utilized the capable machine learning library of python, scikit-learn or sklearn to partition this dataset into Test and Train datasets. Using its model selection approach, or Supervised Learning, to generate testing data by selecting random values from the available dataset for model prediction.

**Step 9:** Evaluate the Model: This step "measures" the model's objective performance using a metric or a set of metrics. Test the model using data that hasn't been seen before. This hidden data is intended to be indicative of model performance in the real world, yet it nevertheless aids in model tuning (as opposed to test data, which does not). Is this a good train/evaluation split? Depending on the domain, data availability, dataset specifics, and other factors, 80/20, 70/30, or comparable ratios may be used.

**Phase 10:** Parameter Tuning: This step is about hyperparameter tuning, which is more of a "artform" than a science. Model parameters that are fine-tuned boost performance. Number of training steps, learning rate, initialization values and distribution, and so on are examples of simple model hyperparameters.

**Making Predictions (Step 11):** More (test set) data that have been kept from the model until now (and for which class labels are available) is used to test the model, providing a better estimate of how the model will behave in the actual world.

## 3.2. Classification Algorithms Used

### 3.2.1. Logistic Regression Technique

For 'Classification' problems, Logistic Regression is a 'Statistical Learning' approach that falls under the genre of 'Supervised' Machine Learning (ML) methods. It has a great reputation, notably in the financial sector, for its extraordinary ability to find defaulters during the last two decades. A basic use pattern for Logistic Regression and other common Linear Classifiers is shown below.

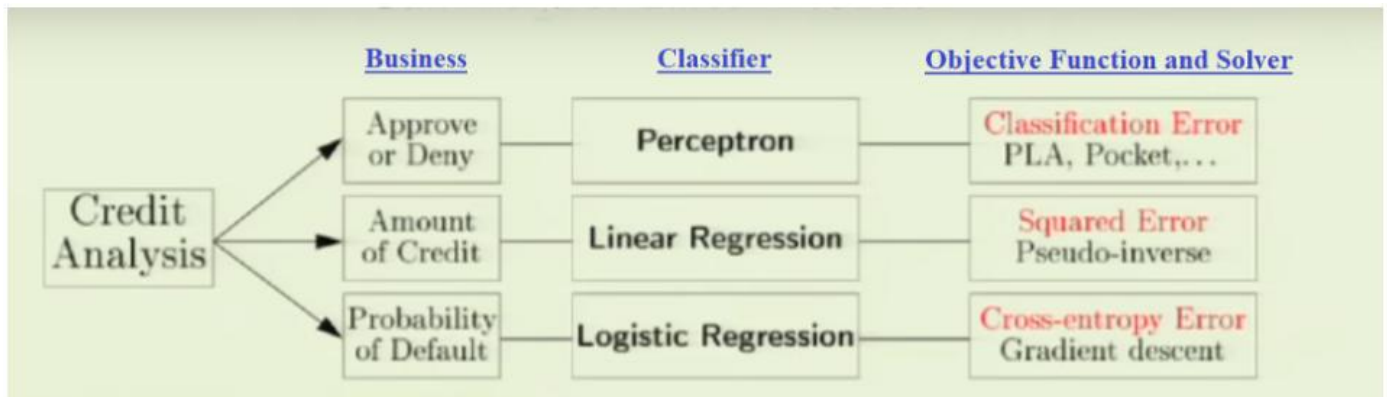


Fig 3.1 :Linear Classifiers and their Usage

A paradox develops when we declare that a classifier with the name 'Regression' is being used for classification, yet this is why Logistic Regression is so magical: it produces discrete binary outputs using a linear regression equation. Yes, it's included in the 'Discriminative Models' subgroup[1] of machine learning approaches such as Support Vector Machines and Perceptrons, which all employ linear equations as a building block and aim to optimise the quality of output on a training set.

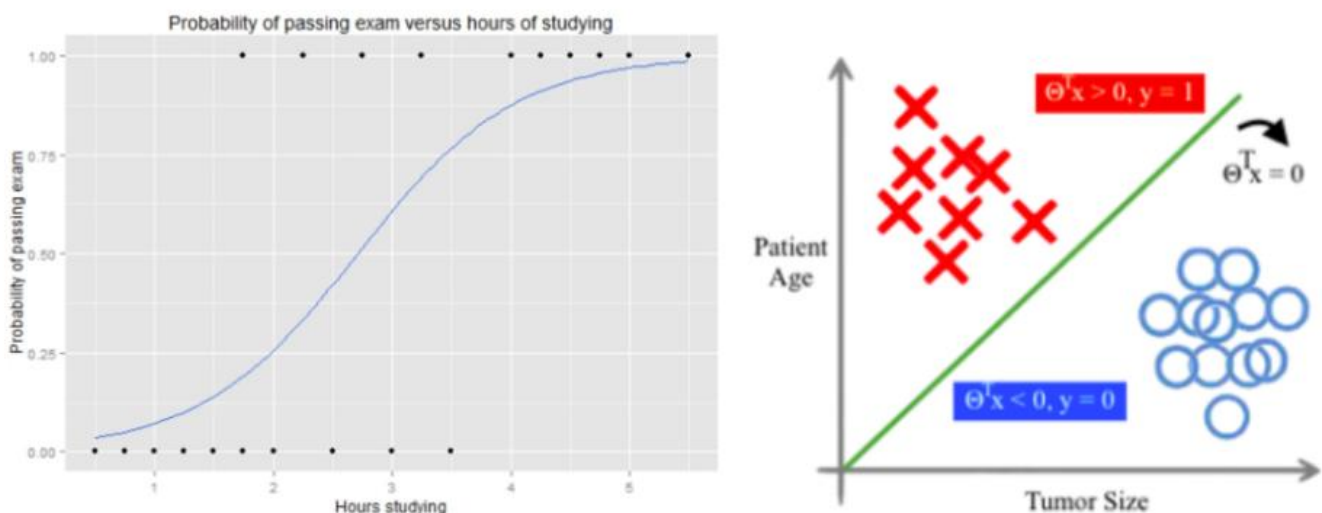


Fig 3.2 :A Journey from Decision Function to Decision Boundary

### 3.2.2. K-Nearest Neighbors Classifier

Similar items, according to the KNN algorithm, are near together. To put it another way, goods that are connected are close to one another.

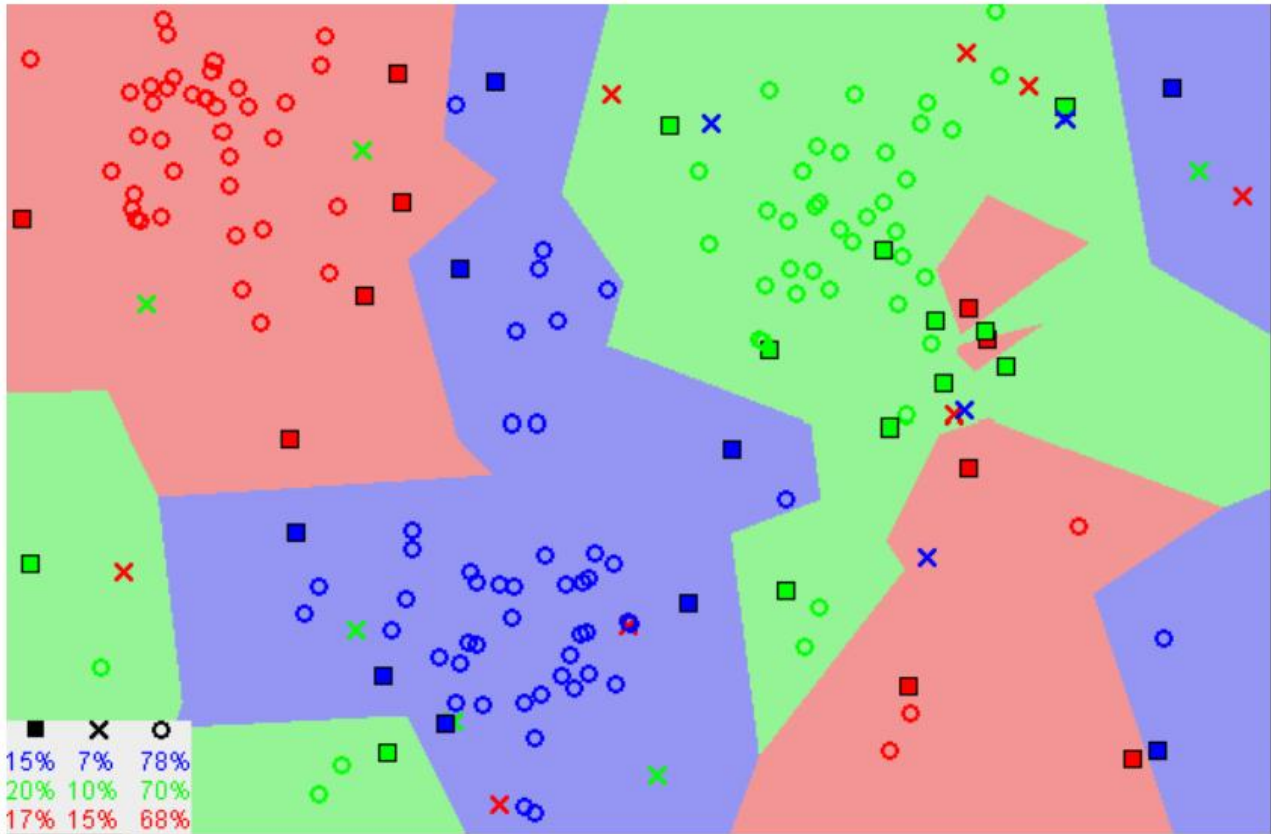


Fig 3.3 :Image showing how similar data points typically exist close to each other

You'll notice that similar data points are typically clustered together in the graph above. In order for the KNN algorithm to work, this assumption must be valid. The concept of similarity (also known as distance, proximity, or closeness) is combined with some basic mathematics, such as computing the distance between points on a graph, in KNN.

It's crucial to understand how we compute the distance between points on a graph before continuing on. Go back and read "Distance Between 2 Points" in its entirety if you're unfamiliar with or need a refresher on how to perform this calculation.

There are several techniques for calculating distance, and one approach may be selected based on the job at hand. A frequent and well-known alternative is the straight-line distance (also known as the Euclidean distance).

### 3.2.3. Support Vector Machine Classifier

The Support Vector Machine (SVM) Classification is similar to the Support Vector Regression (SVR) that I previously explained. The hyperplane in SVM refers to the line that divides the classes. Support Vectors are the closest data points on each side of the hyperplane that are utilised to plot the boundary line.

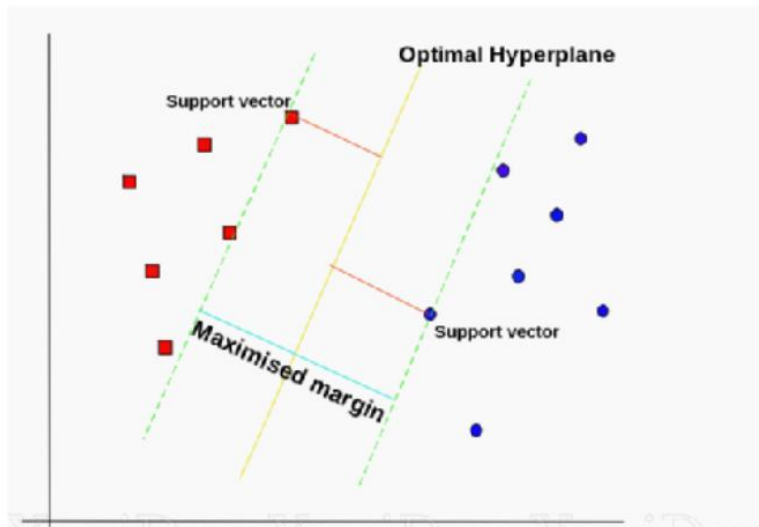


Fig 3.4 :SVM Generalized

SVM Classification data might be linear or non-linear. Different kernels can be chosen in an SVM Classifier.

For a linear dataset, we may set the kernel to 'linear.'

Non-linear datasets, on the other hand, have two kernels: 'rbf' and 'polynomial.'

The data has been transferred to a higher dimension, making it easier to design the hyperplane. Then it's reduced down to the tiniest size imaginable.

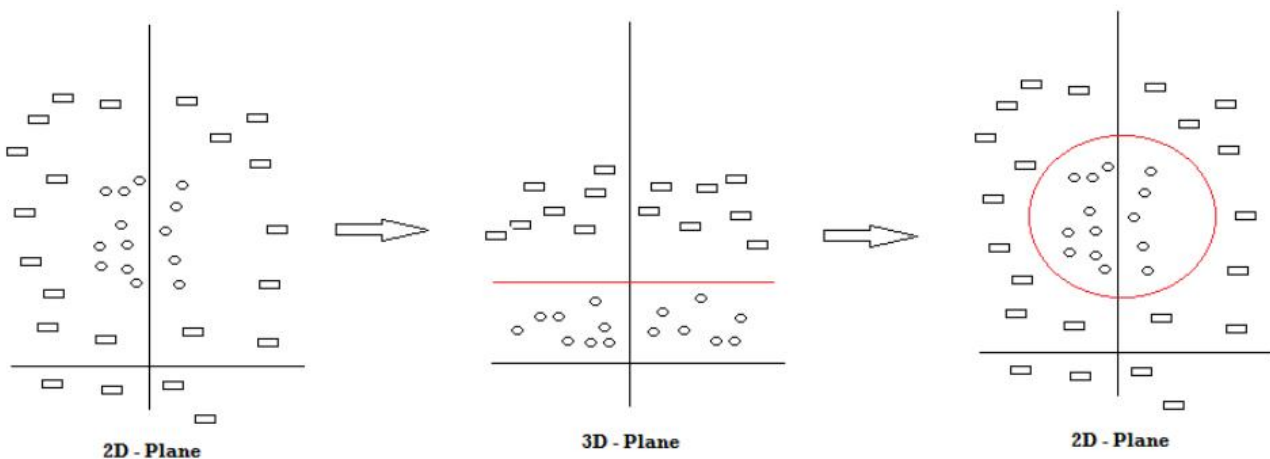


Fig 3.5 :SVM Mechanism

We can see that there are two types of forms in the diagram above: rectangles and circles. We move the data points to a higher dimension (3D Plane) and then build the hyperplane because drawing an SVM line in the 2D Plane is tricky. The SVM Classifier is then drawn in red and brought back to the original plane.

In this approach, the SVM Classifier can categorise a data point from a dataset into which class it belongs. Let's try to tackle a real-world problem with this algorithm.

### 3.2.4. Naive Bayes Classifier

A Foolish Person The Bayes classifier is a probabilistic machine learning model for classification. The Bayes theorem lies at the heart of the classifier.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig 3.6 :Bayes Theorem Equation

Using Bayes' theorem, we may determine the odds of A occurring if B has previously occurred. The hypothesis is A, and the proof is B. In this situation, the predictors/features are supposed to be independent. To put it another way, the presence of one quality has no influence on the presence of the other. As a result, it's dubbed naive.

Naive Bayes algorithms are used in sentiment analysis, spam filtering, recommendation systems, and other applications. They're quick and easy to set up, but the need that predictors be independent is a huge disadvantage. In real-life circumstances, the predictors are frequently dependent, limiting the classifier's usefulness.

Naive Bayes algorithms are used in sentiment analysis, spam filtering, recommendation systems, and other applications. They're quick and easy to set up, but the need that predictors be independent is a huge disadvantage. In real-life circumstances, the predictors are frequently dependent, limiting the classifier's usefulness.

### 3.2.5. Decision Tree Classifier

The Decision Tree is a Supervised Machine Learning Algorithm that makes decisions based on a set of rules, much like humans do.

A Machine Learning classification algorithm may be viewed as a tool for making decisions.

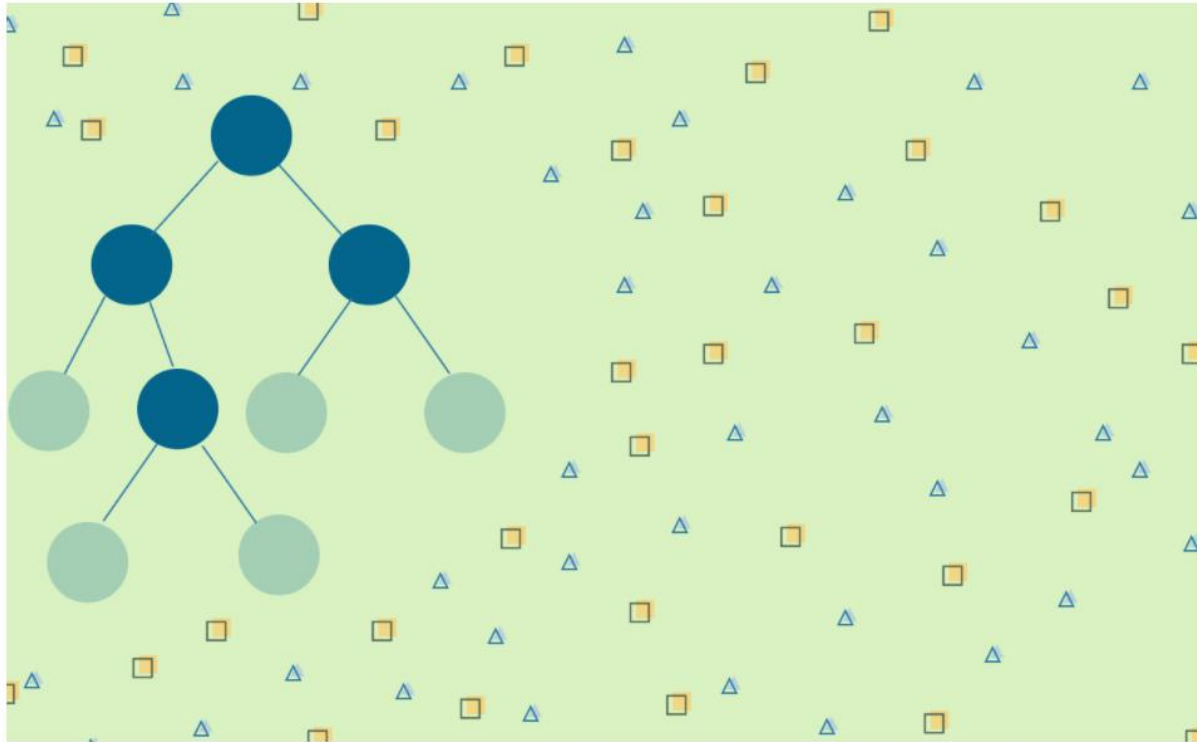


Fig 3.7 :Decision Tree Classifier generalized

Although the model may claim to predict the class of a fresh, never-before-seen input, the algorithm must choose which class to assign behind the scenes.

Some classification algorithms are probabilistic, such as Naive Bayes, but there is also a rule-based method.

On a daily basis, we humans, too, make rule-based judgments.

When planning your next vacation, you employ a rule-based technique. You may select a different place depending on how long you'll be on vacation, your budget, and whether or not you'll be joined by extended relatives.

The answers to these questions will determine the final decision. If you keep narrowing down prospective holiday destinations based on how you answer each question, you might imagine this choosing process as a (decision) tree.

### 3.2.6. Random Forest Classifier

A random forest is made up of a large number of individual decision trees that operate together as an ensemble, as the name suggests. The random forest creates a class prediction for each tree, and the class with the highest votes becomes our model's forecast.

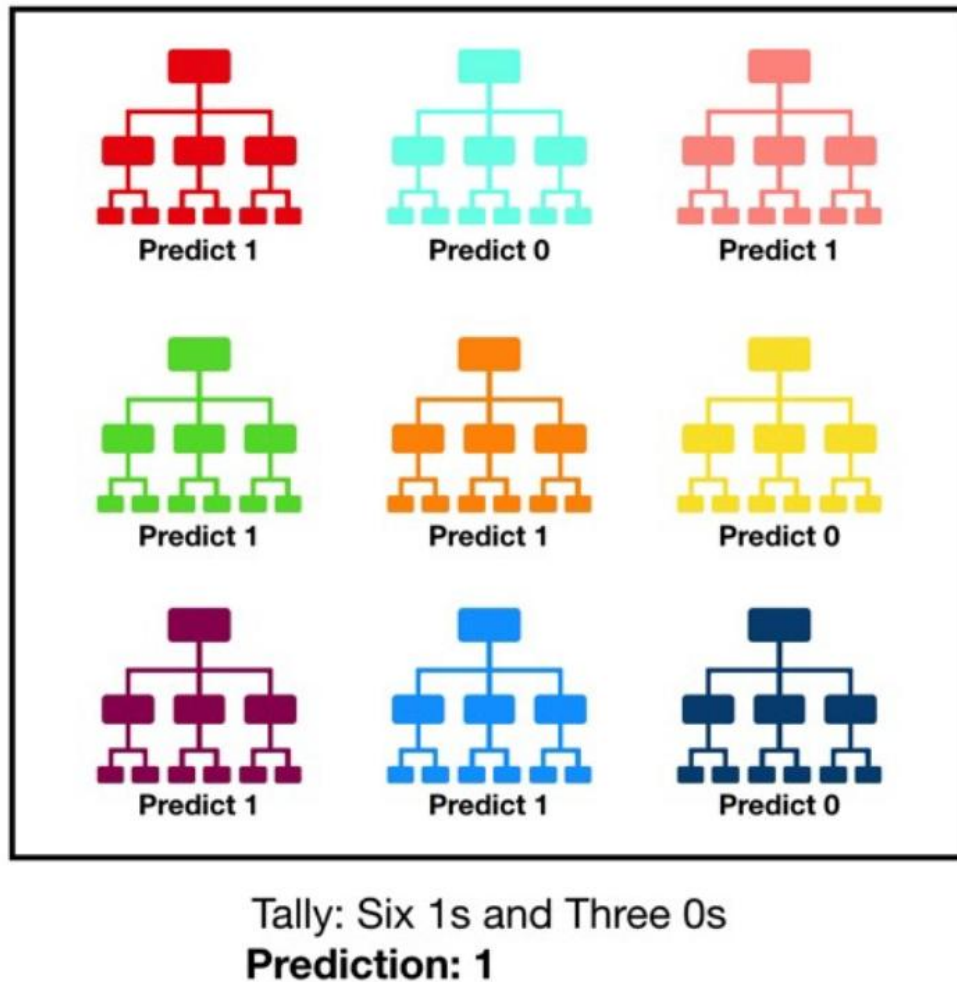


Fig 3.8 :Visualization of a Random Forest Model Making a Prediction

Random forest's underlying idea is the wisdom of crowds, and it's a simple yet powerful one. Some of the reasons why the random forest model works so well in data science are as follows:

A large number of substantially uncorrelated models (trees) functioning as a committee will outperform any of the individual constituent models.

The poor correlation between models is the key.

Uncorrelated models can provide ensemble forecasts that are more accurate than any of the individual projections, just like low-correlation investments (such as stocks and bonds) combine to form a portfolio that is bigger than the sum of its parts. The trees defend each other from their own errors, which explains this wonderful effect.

While some trees will be inaccurate, many others will be correct, allowing the trees to travel in the same direction as the rest of the forest. As a result, the following requirements must be followed in order for random forest to be successful:

To outperform random guessing, models generated using our features must have some true signal, and individual tree projections (and consequently mistakes) must have minimal correlations with one another.

### **3.3. Results Algorithms Used**

**Case A :** Suicidal - Logistic Regression

**Case B :** Depression - Naive Bayes

**Case C :** Depression Severity - Support Vector Machine

### **3.4. Feature Selection**

#### **3.4.1. Feature selection using Extra Tree Classifier**

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning approach that generates a classification result by combining the results of numerous de-correlated decision trees gathered in a "forest." With the exception of how the decision trees in the forest are created, it is theoretically identical to a Random Forest Classifier.

The Decision Trees in the Extra Trees Forest are all based on the original training sample. The tree is then given a random sample of  $k$  features from the feature set at each test node, from which it must select the best feature to partition the data according to specified mathematical criteria (typically the Gini Index).

This random sample of characteristics is used to generate several de-correlated decision trees.

To perform feature selection using the above forest structure, the normalised total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the decision of feature of split) is computed for each feature during the construction of the forest. This number is known as the Gini Importance of the feature. To carry out feature selection, each feature is rated by Gini Importance in descending order, and the user chooses the top  $k$  features depending on his or her preferences.



### 3.4.1.1. Predicting whether the person given in the data is on the verge of attempting suicide or not

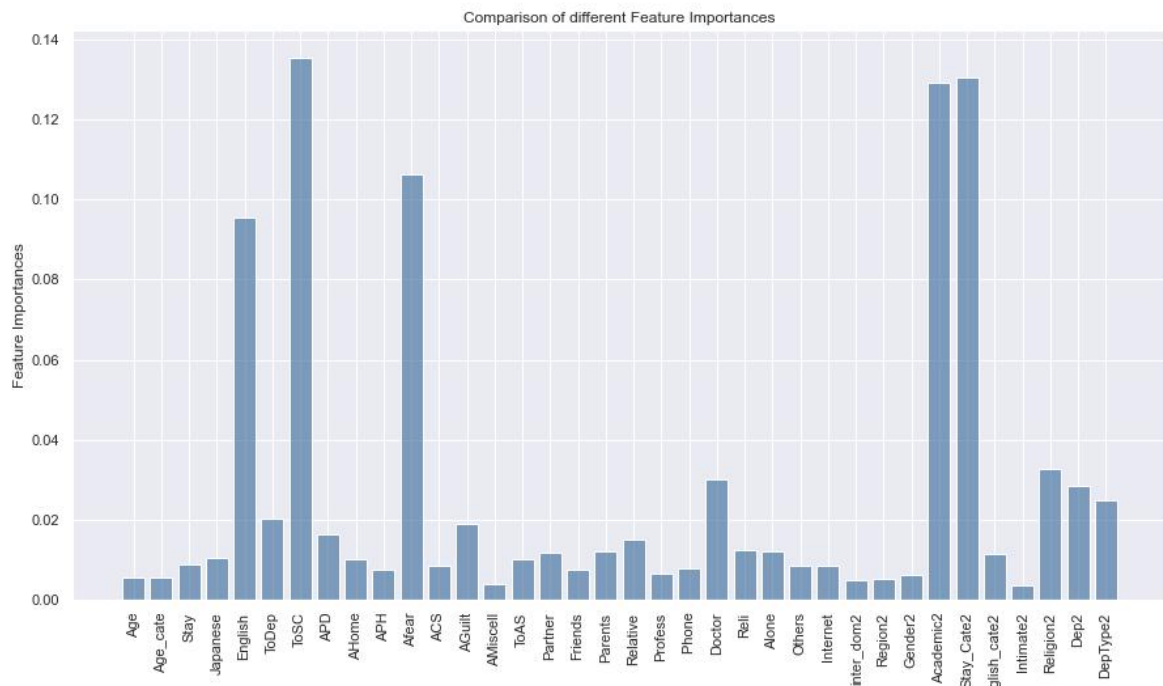


Fig 3.9 : Feature selection results for Case A.

Top 5 Feature Selected	
English	
ToSC	
Afear	
Academic2	
Stay_Cate2	

### 3.4.1.2. Predicting whether a person is affected by depression or not

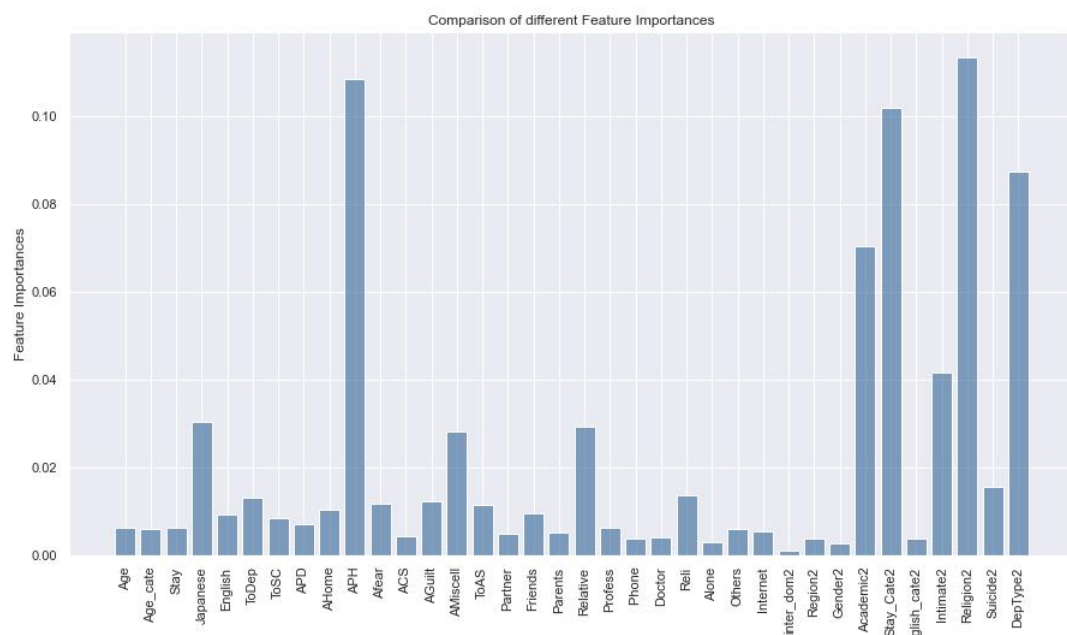


Fig 3.10 : Feature selection results for Case B.

Top 5 Feature Selected
APH
Religion2
DepType2
Academic2
Stay_Cate2

### 3.4.1.3. Predicting the Depression Severity

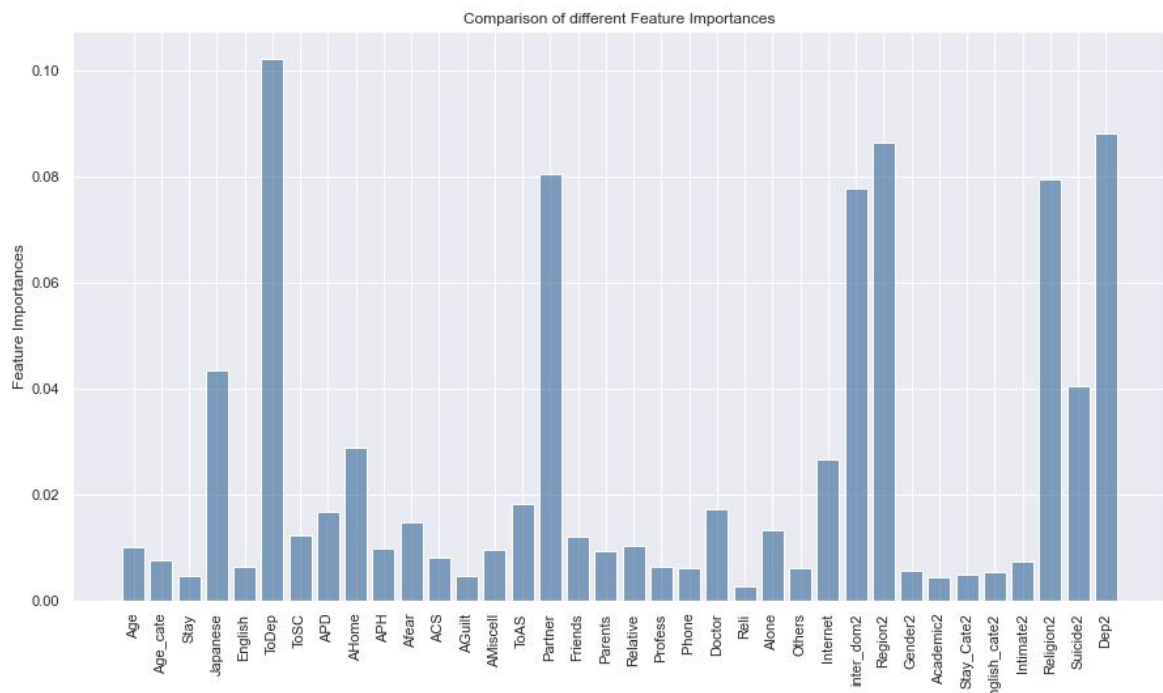


Fig 3.11 : Feature selection results for Case B.

Top 5 Feature Selected
ToDep
Partner
Region2
Religion2
Dep2

### 3.4.2. Heat Map

Colored maps that represent data in a two-dimensional manner are known as heat maps. The colour maps employ hue, saturation, or brightness to produce colour variation in order to portray a variety of features. This colour fluctuation gives readers a visual indication of the magnitude of numerical numbers. Heat Maps replace numbers with colours because the human brain interprets visuals better than numbers, text, or other written data. Because humans are visual learners, visualising data in whatever format makes greater sense. Heat maps are easy-to-understand visual representations of data. As a result, visualisation techniques like Heat Maps have become increasingly popular.

Heat maps may be used to show patterns, variation, and even anomalies, as well as represent the density and intensity of variables. Heatmaps are used to show the relationships between variables. These variables are shown on both axes. We seek for patterns by watching the colour change in the cell. It only accepts numeric input and displays it on a grid, with different colours denoting different data values.

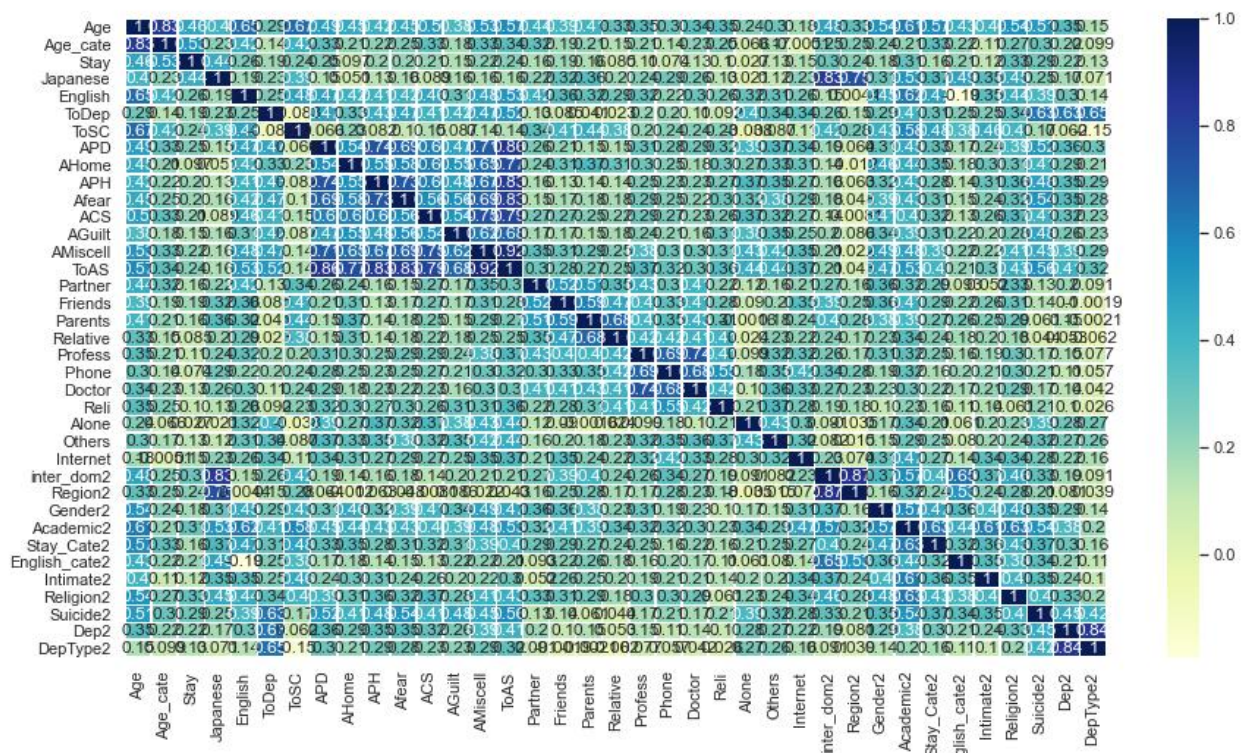


Fig 3.12 : Heat Map

### 3.4.3. Chi Square Test

The chi-square test is used in statistics to assess if two occurrences are independent. From the data of two variables, we may extract the observed count O and predicted count E. The Chi-Square test measures the difference between anticipated and observed counts E and O.

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c$  = degrees of freedom

$O$  = observed value(s)

$E$  = expected value(s)

Fig 3.13 : Formula for chi square test

Consider the following scenario: we need to figure out how the independent category feature (predictor) and the dependent category feature are related (response). When choosing features, we want to choose features that are significantly reliant on the reaction.

When two characteristics are independent, the observed count is close to the predicted count, hence the Chi-Square value is lower. Because the Chi-Square value is so high, the independence hypothesis must be wrong. Simply defined, the greater the Chi-Square value, the more the feature is dependent on the response, and it may be used to train models.

The stages for the Chi-Square Test are outlined below with an example:

Let's take a look at a data set in which we're trying to figure out why clients are quitting a bank. Let's put two variables through a Chi-Square test. Exited is a response that indicates whether a client is leaving the bank, and Gender is a prediction that determines whether a consumer is male or female. In this test, we'll determine if there's a relationship between Gender and Exited.

- Define hypothesis.
- Make a table for contingencies.
- Determine the expected results.
- The Chi-Square statistic should be calculated.
- Accept the Null Hypothesis or reject it.

### 3.5. Analyzing the Output

The tests are run on a pre-processed dataset, and the methodologies stated above are explored and applied. The confusion matrix is used to calculate the above-mentioned performance indicators. The Confusion Matrix describes the model's performance.

#### 3.5.1. Confusion Matrix

A confusion matrix is a table that summarises the results of classification problem prediction. Count values are used to sum and break down the number of correct and incorrect predictions by class. This is the key to the confusion matrix. It not only notifies you about the errors produced by your classifier, but also about the sorts of errors made. The disadvantage of depending only on classification accuracy is addressed in this analysis.

$$\text{Precision} = (TP + FP) / (TP)$$

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

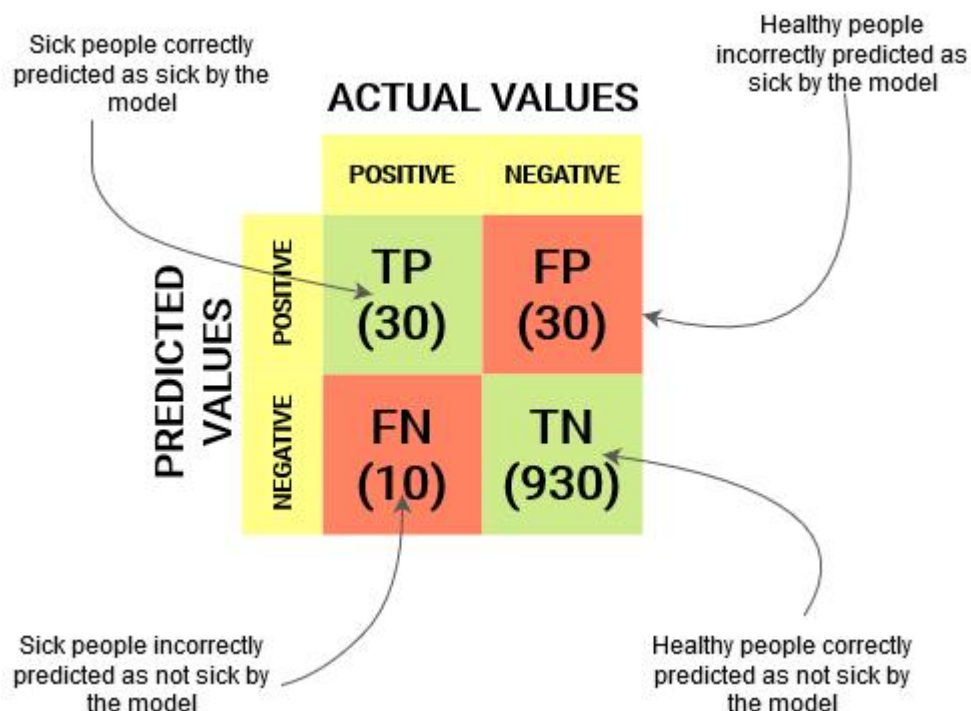
$$F (\text{Precision} + \text{Recall}) = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

True positive (TP) means the patient has the ailment and the test findings are positive.

FP - False positive: the test results are positive despite the fact that the patient does not have the ailment.

TN - The patient does not have the disease, and the tests have come back negative.

False negative (FN): despite the fact that the patient has the disease, the test results are negative.





**Recall** is defined as the proportion of true positive examples divided by the total number of positive and false negative examples. The number of real positive cases that our model was able to correctly predict is referred to as recall. Recall is a useful measure when False Negative outnumbers False Positive. In medical situations, recall is essential since, regardless of whether we raise a false alert, real positive instances should not go ignored!

**Precision** : Precision refers to how many of the correctly predicted instances turned out to be positive. It's the total number of true positives + false positives divided by the number of real positives. Precision is a useful statistic when False Positives are more of a concern than False Negatives.

Note: A better statistic would be recall since we don't want to discharge a medically impaired individual into the general public by accident.

**Accuracy** is the most fundamental. The total number of correct predictions in the full dataset is specified. The equation is formed by dividing true positive and true negative cases by true positive, false positive, true negative, and false negative examples.

**The F1-score** gives a composite view of Precision and Recall since it is a harmonic mean of these two measurements. It achieves its pinnacle when Precision equals Recall.

There is, however, a catch. The F1-score is difficult to understand. That is, we have no way of knowing whether our classifier maximises precision or recall. As a consequence, we blend it with other assessment indicators to provide a more comprehensive picture of the outcome.

**Importance of Confusion Matrix** : Because our classifier is attempting to predict if a patient has a given ailment based on the symptoms (features) that are fed into it, the Confusion Matrix is critical. The patient either has or does not have the illness since this is a binary classification problem. On the left-hand side of the confusion matrix, the classifier's predicted class is presented. Meanwhile, the samples' real class designations are maintained in the top row of the matrix. Implementing the Confusion Matrix

- a) First, we'll need to import the necessary Python modules. The logistic regression classifier from Sklearn, as well as the train/test split tool and the confusion matrix measure, were all employed. Numpy and Matplotlib will also be used.
- b) The data is transformed into a type that our classifier can understand using Numpy, and the data is visualized using Matplotlib.

- c) As well as characteristics and labels, our data must now be separated into a training set and a testing site. The "X"s stand for features, whereas the "Y"s stand for labels.
- d) Now we'll create a classifier instance. The logistic regression classifier, which we earlier imported, was employed.
- e) After building a classifier instance, we can simply use the "fit" command to fit and train it on the data. We may now feed pictures for categorization to the fitted classifier. We may either feed it a single image at a time or slice the data set into many photographs.
- f) Now, without utilising the "X test" variable, make predictions using the entire test data set. We'll also keep the predictions in a variable called "predictions" so that we can evaluate the classifier's performance using them.
- g) All that remains was to assess the performance of our classifier. To gain a fast idea of the classifier's accuracy, we simply used the "score" command on the logistic regression object.
- h) To obtain our classifier's confusion matrix, we must first create an instance of the Sklearn confusion matrix and pass it the needed arguments: the true values and our predictions.
- i) In Sklearn, we created a confusion matrix in this manner. There are various methods for generating a confusion matrix in Python, such as utilising the Seaborn module, but this is one of the most straightforward.

## Conclusion

A confusion matrix is an effective tool for predictive analysis since it allows you to compare expected and actual values. It will take some practise to master the skill of understanding a confusion matrix, but once you have, it will be an invaluable tool in your data scientist toolbox.

### 3.6. Observations(Precision,recall,f1,accuracy scores) in Table form.

We have put together the precision, recall, f1, support and all the accuracy scores along with the confusion matrices in table form as for each of the three cases:

#### 3.6.1. Predicting whether the person given in the data is on the verge of attempting suicide or not

Logistic Regression					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0]
Class 0	0.91	0.95	0.93	42	
Class 1	0.75	0.6	0.67	10	[ 0 40 2]
Accuracy			0.90	58	
Micro avg	0.89	0.85	0.87	58	[ 0 4 6]]
Weighted avg	0.89	0.90	0.89	58	
Accuracy Score	0.896551724137931				

Table 3.1 : Logistic Regression accuracy results.

KNeighbors Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0]
Class 0	0.88	0.90	0.89	42	
Class 1	0.56	0.50	0.53	10	[ 0 38 4]
Accuracy			0.84	58	
Micro avg	0.81	0.8	0.81	58	[ 0 5 5]]
Weighted avg	0.84	0.84	0.84	58	
Accuracy Score	0.844827586206896				

Table 3.2 : KNN Classifier accuracy results.

Decision Tree Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	0.83	0.91	6	[[ 5 0 1]
Class 0	0.85	0.95	0.90	42	
Class 1	0.50	0.30	0.37	10	[ 0 40 2]
Accuracy			0.83	58	
Micro avg	0.78	0.70	0.73	58	[ 0 7 3]]
Weighted avg	0.81	0.83	0.88	58	
Accuracy Score	0.827586206896551				

Table 3.3 : Decision Tree Classifier accuracy results.



Support Vector Machine					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0]
Class 0	0.91	0.95	0.93	42	
Class 1	0.75	0.60	0.67	10	[ 0 40 2]
Accuracy			0.90	58	
Micro avg	0.89	0.85	0.87	58	[ 0 4 6]]
Weighted avg	0.89	0.90	0.89	58	
Accuracy Score	0.896551724137931				

Table 3.4 : SVM Classifier accuracy results.

Naive Bayes Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0]
Class 0	0.91	0.93	0.92	42	
Class 1	0.67	0.60	0.63	10	[ 0 39 3]
Accuracy			0.88	58	
Micro avg	0.86	0.84	0.85	58	[ 0 4 6]]
Weighted avg	0.88	88.00	0.88	58	
Accuracy Score	0.879310344827586				

Table 3.5 : Naive Bayes Classifier accuracy results.

Random Forest Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0]
Class 0	0.85	0.98	0.91	42	
Class 1	0.75	0.30	0.43	10	[ 0 41 1]
Accuracy			0.86	58	
Micro avg	0.87	0.76	0.78	58	[ 0 7 3]]
Weighted avg	0.85	0.86	0.84	58	
Accuracy Score	0.862068965517241				

Table 3.6 : Random Forest Classifier accuracy results.

### 3.6.2. Predicting whether a person is affected by depression or not

Logistic Regression					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	0.80	0.89	5	[[ 4 0 1 0]
Class 0	0.91	0.97	0.94	33	
Class 1	0.84	0.84	0.84	19	[ 0 32 1 0]
Class 3	0.00	0.00	0.00	1	
Accuracy			0.90	58	[ 0 3 16 0]
Micro avg	0.69	0.65	0.67	58	
Weighted avg	0.88	0.90	0.89	58	[ 0 0 1 0]]
Accuracy Score	0.896551724137931				

Table 3.7 : Logistic Regression accuracy results.

KNeighbors Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	0.83	1.00	0.91	5	[[ 5 0 0 0]
Class 0	0.7	0.91	0.79	33	
Class 1	0.67	0.32	0.43	19	[ 0 30 3 0]
Class 3	0.00	0.00	0.00	1	
Accuracy			0.71	58	[ 0 13 6 0]
Micro avg	0.55	0.56	0.53	58	
Weighted avg	0.69	0.71	0.67	58	[ 1 0 0 0]]
Accuracy Score	0.706896551724137				

Table 3.8 : KNN Classifier accuracy results.

Decision Tree Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	0.83	0.91	6	[[ 5 0 1]
Class 0	0.85	0.95	0.90	42	
Class 1	0.50	0.30	0.37	10	[ 0 40 2]
Accuracy			0.83	58	
Micro avg	0.78	0.70	0.73	58	[ 0 7 3]]
Weighted avg	0.81	0.83	0.81	58	
Accuracy Score	0.8275862068965570				

Table 3.9 : Decision Tree Classifier accuracy results.

Support Vector Machine					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	0.80	0.89	5	[[ 4 0 0 1 0]
Class 0	1.00	1.00	1.00	33	
Class 1	1.00	1.00	1.00	19	[ 0 33 0 0 0]
Class 2	0.00	0.00	0.00	0	[ 0 0 19 0 0]
Class 3	0.00	0.00	0.00	1	
Accuracy			0.97	58	[ 0 0 0 0 0]
Micro avg	0.60	0.56	0.58	58	
Weighted avg	0.98	0.97	0.97	58	[ 0 0 0 1 0]]
Accuracy Score	0.965517241379310				

Table 3.10 : SVM Classifier accuracy results.

Naive Bayes Machine					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	0.80	0.89	5	[[ 4 0 1 0]
Class 0	1.00	1.00	1.00	33	
Class 1	0.9	1.00	0.95	19	[ 0 32 1 0]
Class 3	0.00	0.00	0.00	1	[ 0 3 16 0]
Accuracy			0.97	58	
Micro avg	0.73	0.70	0.71	58	[ 0 0 1 0]]
Weighted avg	0.95	0.97	0.96	58	
Accuracy Score	0.965517241379310				

Table 3.11 : Naive Bayes Classifier accuracy results.

Random Forest Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0]
Class 0	0.85	0.98	0.91	42	
Class 1	0.75	0.30	0.43	10	[ 0 41 1]
Accuracy			0.86	58	[ 0 7 3]]
Micro avg	0.87	0.76	0.78	58	
Weighted avg	0.85	0.86	0.84	58	
Accuracy Score	0.862068965517241				

Table 3.12 : Random Forest Classifier accuracy results.



### 3.6.3. Predicting the Depression Severity

Logistic Regression					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	0.80	1.00	0.89	4	[[ 4 0 0 0 0 0]
Class 0	0.92	1.00	0.96	33	
Class 1	0.59	0.77	0.67	13	[ 0 33 0 0 0 0]
Class 2	0.00	0.00	0.00	6	
Class 4	0.00	0.00	0.00	1	[ 0 3 10 0 0 0]
Class 5	0.00	0.00	0.00	1	
Accuracy			0.81	58	[ 0 0 6 0 0 0]
Micro avg	0.38	0.46	0.42	58	
Weighted avg	0.71	0.81	0.75	58	[ 0 0 1 0 0 0]
Accuracy Score	0.810344827586206				[ 1 0 0 0 0 0]]

Table 3.13 : Logistic Regression accuracy results.

KNeighbors Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	0.80	1.00	0.89	4	[[ 4 0 0 0 0 0]
Class 0	1.00	1.00	1.00	33	
Class 1	0.68	1.00	0.81	13	[ 0 33 0 0 0 0]
Class 2	0.00	0.00	0.00	6	
Class 4	0.00	0.00	0.00	1	[ 0 0 13 0 0 0]
Class 5	0.00	0.00	0.00	1	
Accuracy			0.86	58	[ 0 0 6 0 0 0]
Micro avg	0.41	0.50	0.45	58	
Weighted avg	0.78	0.86	0.81	58	[ 0 0 0 1 0 0]
Accuracy Score	0.862068965517241				[ 1 0 0 0 0 0]]

Table 3.14 : KNN Classifier accuracy results.

Decision Tree Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	0.83	0.91	6	[[ 5 0 1]
Class 0	0.85	0.93	0.89	42	
Class 1	0.43	0.30	0.35	10	[ 0 39 3]
Accuracy			0.81	58	
Micro avg	0.76	0.69	0.72	58	[ 0 7 3]]
Weighted avg	0.79	0.81	0.80	58	
Accuracy Score	0.810344827586206				

Table 3.15 :Decision Tree Classifier accuracy results.

Naive Bayes Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	0.80	1.00	0.89	4	[[ 4 0 0 0 0 0 ]]
Class 0	1.00	1.00	1.00	33	
Class 1	0.68	1.00	0.81	13	
Class 2	0.00	0.00	0.00	6	[ 0 33 0 0 0 0 ]
Class 4	0.00	0.00	0.00	1	[ 0 0 13 0 0 0 ]
Class 5	0.00	0.00	0.00	1	
Accuracy			0.86	58	[ 0 0 6 0 0 0 ]
Micro avg	0.41	0.50	0.45	58	[ 0 0 0 1 0 0 ]
Weighted avg	0.78	0.86	0.81	58	[ 0 0 0 1 0 0 ]
Accuracy Score	0.862068965517241				[ 1 0 0 0 0 0 ]]

Table 3.16 : Naive Bayes Classifier accuracy results.

Support Vector Machine					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	0.80	1.00	0.89	4	[[ 4 0 0 0 0 0 0 ]]
Class 0	1.00	1.00	1.00	33	
Class 1	0.92	0.92	0.92	13	
Class 2	0.83	0.83	0.83	6	[ 0 0 12 1 0 0 0 ]
Class 3	0.00	0.00	0.00	0	
Class 4	0.00	0.00	0.00	1	[ 0 0 1 5 0 0 0 ]
Class 5	0.00	0.00	0.00	1	
Accuracy			0.93	58	[ 0 0 0 0 0 0 0 ]
Micro avg	0.51	0.54	0.52	58	[ 0 0 0 0 1 0 0 ]
Weighted avg	0.92	0.93	0.92	58	[ 0 0 0 0 1 0 0 ]
Accuracy Score	0.931034482758620				[ 1 0 0 0 0 0 0 ]]

Table 3.17 : SVM Classifier accuracy results.

Random Forest Classifier					
	Precision	Recall	f1-score	Support	Confusion Matrix
Class -1	1.00	1.00	1.00	6	[[ 6 0 0 ]]
Class 0	0.85	0.98	0.91	42	
Class 1	0.75	0.30	0.43	10	
Accuracy			0.86	58	[ 0 41 1 ]
Micro avg	0.87	0.76	0.78	58	[ 0 7 3 ]]
Weighted avg	0.85	0.86	0.84	58	
Accuracy Score	0.862068965517241				

Table 3.18 : Random Forest Classifier accuracy results.

## LITERATURE REVIEW

**1. Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. Journal of psychiatric research, 47(3), 391-400.**

Despite being a socially advantaged demographic, this study found substantial evidence that university students are more likely to suffer from depression. The major goals of this review are to find research that report on depression rates among university students and to test the hypothesis that depression rates are rising among undergraduate university students.

### METHODOLOGY:

1. Studies on the prevalence of depression among university students published between 1990 and 2010 were obtained by searching PubMed, PsycINFO, BioMed Central, and Medline.
2. Search terms included depression, depressive symptoms, depressive disorders, prevalence, university students, college students, undergraduate students, adolescents, and/or young adults. Each study was given a quality rating.
3. Because sample size and response rate are so important in any prevalence study, great care should be taken in determining and reporting them.

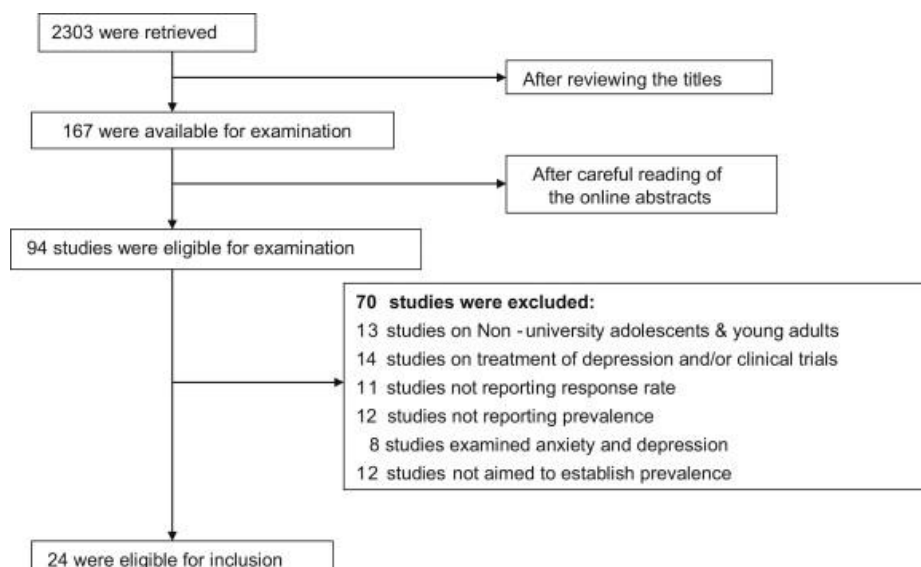


Fig 4.1 : Methodology flowchart for Journal of psychiatric research, 47(3), 391-400.

## **PROS:**

- This research is made up of multiple nested cross-sectional studies with general population samples. As a result, there is a well-validated and reliable method for screening depression among university students.
- The researchers also discovered a wide range of inclusion and exclusion criteria, as well as techniques for diagnosing depression and determining its severity.

## **CONS:**

- The main limitation was the possibility of missing studies that did not directly report on depressive prevalence (i.e. studies examining the prevalence of general distress and using measures that screen for depression as one of the elements of general distress, such as the General Health Questionnaire of the Symptom Checklist (SCL-90)).
- The co-morbidity of anxiety and depression may cause the prevalence rates in the studies to be overstated.
- The average prevalence of depression (30.6 percent) in the current study might have been overstated by include certain studies that only reported rates of major depressive disorder rather than milder depressed states.
- Only a few studies were included in this evaluation since several studies confirmed the frequency of depression but did not provide a response rate. This is crucial because, because non-respondents and respondents may differ in ways other than their desire to engage in a survey, the study is less valid (for both external and internal validity) (Denscombe, 2008, 2009).

## **CONCLUSION:**

The data was extensively analysed, and the quality evaluation approach for epidemiological prevalence studies proposed by Parker and colleagues (Parker et al., 2008) was used. Each item received one score for each of the following quality indicators:

- (1) the target population was clearly defined,
- (2) complete, random, or consecutive recruitment,
- (3) the targeted sample is representative or the report presents evidence that the results can be generalised to the general undergraduate population,
- (4) the response rate was equal to or greater than 70%,
- (5) the scale used is a validated measure of depression with valid cut-offs for depression classification, and
- (6) the sample size is adequate with a minimum sample size of 100. Because the larger the sample, the more precise the results are, the last two quality criteria were included (Strachan, 1997).

Furthermore, establishing the trustworthiness of prevalence research data requires the use of CI and SE. Either a CI or a SE should be calculated and included in the study's findings. According to the research, university students had much greater rates of depression than the general population. The inclusion and exclusion criteria were fulfilled by twenty-four papers. The stated prevalence percentages varied from 10% to 85%, with a weighted average prevalence of 30.6 percent.

## **2. Sandhu, D. S., & Asrabadi, B. R. (1994). Development of an acculturative stress scale for international students: Preliminary findings. *Psychological reports*, 75(1), 435-448.**

This research presents the creation and testing of a new 36-item Likert scale to evaluate foreign students' acculturative stress, which incorporates important contributing elements such as felt discrimination, homesickness, fear, guilt, perceived hatred, and stress due to change (culture shock). The psychometric characteristics of this instrument, as well as the implications for its use by mental health practitioners, are also discussed in this work.



## **METHODOLOGY:**

- Correlation and factor analysis were among the approaches employed to examine the data.
- To execute the essential computations, the SPSS Release 3.0 for UNISYS big computers was employed.
- The anti-image correlation (the negative of the partial correlation coefficient) proved the validity of component analysis, and Bartlett's test of sphericity was used to verify the hypothesis that the population correlation matrix was an identity matrix.
- The sphericity test statistic yielded a value of 3678.84 and a significance level of p.00001. Despite the fact that the test statistic was big enough to overcome the lack of normality, it was based on the assumption that data were drawn from a multivariate normal population.

## **PROS:**

- This study examines the search of knowledge beyond indigenous borders.
- This measure may be beneficial for practitioners in identifying and assessing foreign students' acculturative stress and devising particular techniques to assist them.
- Researchers might also use this measure to compare acculturative stress experiences among international students of diverse ethnic groups, and utilise that data to assess the effectiveness of counselling treatments.
- The scale also quantifies acculturative stress, perhaps facilitating further empirical study.

## **CONS:**

- Apparently, research on overseas students' psychological issues is isolated, sporadic, inconsistent, diversified, and desultory in character.
- The majority of overseas students' psychological issues have been hypothesized with little actual evidence to back them up.

## **CONCLUSION:**

Using the main component approach, six components were discovered, accounting for 70.6 percent of the total explained variance. The eigen-values, variance percentages, and cumulative percentages of these variables. Total scores on this scale range from 36 to 180. Individuals with higher scores are enduring more acculturative stress. The scores on six subscales can be determined by aggregating the individual scores on the corresponding items.

**3. Wei, M., Heppner, P. P., Mallen, M. J., Ku, T. Y., Liao, K. Y. H., Wu, T. F. (2007). Acculturative stress, perfectionism, years in the United States, and depression among Chinese international students. *Journal of Counseling Psychology*, 54(4), 385.**

The purpose of this study was to see if maladaptive perfectionism (i.e., a gap between expectations and performance) and the amount of time spent in the United States had an effect on acculturative stress and sorrow. Online questionnaires were completed by 189 Chinese foreign students from China and Taiwan who were attending a midwestern university.

## **METHODOLOGY:**

- The Acculturative Stress Scale for International Students was used to assess acculturative stress (ASSIS; Sandhu & Asrabadi, 1994)
- The Discrepancy subscale of the Almost Perfect Scale—Revised (APS-R; Slaney et al., 2001) was used to assess maladaptive perfectionism.
- The Center for Epidemiological Studies—Depression Scale (CES-D; Radloff, 1977) was used to assess depression.
- We used a multivariate analysis of variance to see if the major measures (acculturative stress, maladaptive perfectionism, and depression) differed by gender, marital status, place of origin, and language version.
- We also looked at whether sex, marital status, country of origin, and different language versions had any interaction effects on the key factors.

- In the analysis, a hierarchical regression was applied.
- To help in the comprehension of the three-way interaction, we additionally evaluated the significant levels for simple interactions and then each of the simple slopes (see Cohen et al., 2003, pp. 290–291). To reduce down the three-way interaction into a more accessible form, Cohen et al. proposed assessing the importance of the basic interaction. When the third predictor has various values or levels, the simple interaction investigates the significance of a two-way interaction between two predictors.

### **PROS:**

- A hierarchical regression revealed that acculturative stress and maladaptive perfectionism had significant main effects on depression, implying that acculturative stress, no significant two-way interactions, and a significant three-way interaction, maladaptive perfectionism, and length of time in the United States interacted to predict depression.
- This study looks at how the amount of time spent in the United States, along with maladaptive perfectionism, helped Chinese foreign students cope with acculturative stress and despair.
- We are confident in the accuracy and completeness of the data we used for data analysis.

### **CONS:**

- Only those who had lived in the United States for a longer amount of time had low levels of maladaptive perfectionism, which helped to mitigate the effects of acculturative stress on depression.
- Because this study only covers students who are interested in or willing to engage in it, the sample size may be biased.
- We observed a significant incidence of worthless surveys after examining the three validity items (55 of 252, 22 percent). The reasons for this might be due to a lack of confidence in the validity items' direction, weariness, rage, and irritation generated by completing a lengthy survey, or simply a mistake made at a specific moment in the survey response.

## **CONCLUSION:**

Acculturative stress, maladaptive perfectionism, and years spent in the United States were shown to account for 49% of depression variation. Acculturative stress and maladaptive perfectionism were revealed to be significant predictors of depression. Years in the United States, on the other hand, failed to predict the onset of the Great Depression. The aggregate two-way interactions did not substantially add extra variation to depression in Step 2 over and beyond the first-order effects. In any of the two-way interactions, there was no statistical significance found. A three-way interaction, however, significantly contributed extra variation in depression in Step 3 beyond the first-order effects and two-way interaction effects. According to Cohen (1992), an  $R^2$  value of .02 indicates a minimal impact magnitude. Acculturative stress, maladaptive perfectionism, and years spent in the United States were all found to be significant predictors of depression. Despite the fact that the interaction term's coefficient looked to be moderate, Cohen et al. (2003, p. 297) highlighted that the effect size for interactions in psychology and social science research tends to be tiny (i.e., squared semipartial or partial correlations of .01–.05 or so).

## **4. Nguyen, M. H., Le, T. T., & Meirmanov, S. (2019). Depression, acculturative stress, and social connectedness among international university students in Japan: a statistical investigation . *Sustainability*, 11(3), 878.**

The goal of this study is to investigate the frequency of depression among domestic and international students at a Japanese international university, as well as its relationship to acculturative stress and social connectivity.

## **METHOD:**

268 students replied to a web-based survey issued to several classes of university students. A nine-item questionnaire from the Patient Health Questionnaire (PHQ-9), the Social Connectedness Scale (SCS), and the Acculturative Stress Scale for International Students (ASSIS) was used in the survey, along with socio-demographic data.

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right)_{depression} &= \ln(Odd Ratios)_{depression} \\
&= \alpha + \beta_{1j}gender_{1j} + \beta_{2j}age_{2j} + \beta_{3j}stay_{3j} + \beta_{4j}Eng_{4j} + \beta_{5j}Jap_{5j} \\
&+ \beta_{6j}partner_{6j} + \beta_{7j}religion_{7j} + e,
\end{aligned}$$

with

$p$ : the probability of being depressed

$\alpha$ : intercept

$\beta$ : coefficient which is the logarithm of Odd Ratios

$j$ : categorical factor of independent variables

$gender, age, etc.$ : independent variables

$e$ : error term.

Fig 4.2 : Accumulative stress scale,SCS,PHQ-9 evaluation formulation.

Based on data obtained from an international university in Japan, the study aims to address the following research questions (RQ1 and RQ2) and hypotheses (H1 and H2) based on the reviewed literature:

**RQ1:** What percentage of domestic and overseas students suffer from depression?

**RQ2:** What are the socioeconomic and demographic factors that predict depression among domestic and overseas students?

**H1:** In both local and overseas students, acculturative stress will be significantly linked to depression. Due to frequent disagreements during the acculturation process, students at an international university were projected to experience higher depression levels.

**H2:** In both local and overseas students, social connectivity will be highly related with depression. Individuals will feel more at ease and confident in social situations if they feel more connected to others, which will help them avoid depression.

Based on data obtained from an international university in Japan, the study aims to address the following research questions (RQ1 and RQ2) and hypotheses (H1 and H2) based on the reviewed literature:

**RQ1:** What percentage of domestic and overseas students suffer from depression?

**RQ2:** What are the socioeconomic and demographic factors that predict depression among domestic and overseas students?

**H1:** In both local and overseas students, acculturative stress will be significantly linked to depression. Due to frequent disagreements during the acculturation process, students at an international university were projected to experience higher depression levels.

**H2:** In both local and overseas students, social connectivity will be highly related with depression. Individuals will feel more at ease and confident in social situations if they feel more connected to others, which will help them avoid depression.

### **PROS:**

- The researchers discovered a link between acculturative stress and depression in both domestic and overseas students in this study.

### **CONS:**

- The team employed sampling in the data collection procedure and had to make minor changes to the model questionnaire.
- Furthermore, the conclusions were based on self-reported data.
- In addition, variable proportions of students from various backgrounds may produce regional bias in the results among the polled international students.

### **CONCLUSION:**

In a cross-sectional questionnaire of 67 domestic students and 201 foreign students, the study found that 29.85 percent of domestic students and 37.81 percent of international students were positive to major depressive illness or other depressive disorder. A variety of socio-demographic factors (age, English competence, and duration of stay) as well as main depression associations are also shown in the data (social connectedness and acculturative stress).

**5. Vuong, Q. H., Vuong, T. T., Ho, T. M., & Nguyen, H. V. (2017). Psychological and socio-economic factors affecting social sustainability through impacts on perceived health care quality and public health: The case of Vietnam. *Sustainability*, 9(8), 1456.**

The impact of psychological and socio-economic variables on patients' views of healthcare quality and public health were investigated in a study of over 2000 patients in Hanoi, Vietnam. Good health communication and marital status, according to the statistics, are two criteria that have the biggest influence on people's positive impressions of healthcare quality. Young, single people, as well as insured people, are harsher on healthcare quality. At the same hand, a higher BMI and a better perception of health-care quality are linked to negative community attitudes toward health. These findings show that, in order to preserve community health as part of social sustainability, the Vietnamese government should focus on infrastructure upgrades, insurance system changes, and personal health care knowledge transmission.

**METHOD:**

This study gathered data on a wide range of socioeconomic and demographic variables. Patients were chosen at random with no discriminating criteria, and roughly 83 percent of them responded (5 out of 6). The interviewer prompted participants to fill out each form correctly, and the questions were simple and easy. The raw data was first imported into Excel and then processed using R. To analyse the influence of demography, society, and psychology on the patient's evaluation of health care service quality, multivariate linear regression was used in conjunction with the general model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

With the stipulation that the sample size of k independent variables  $X_i$  be the same as the sample size of the dependent variable Y.  $X_i$  is a category variable, while Y is a numerical variable. After the data was processed in R, the values of I that indicated the linear influence of  $X_i$  on Y, namely the value of "SerQual" in this study, were produced.

**Table 2.** A few descriptive statistics for continuous variables.

Characteristics	Average	SD	95% CI
Age	29.17	10.09	28.74–29.60
BMI	20.85	2.69	20.73–20.97
Health services quality	3.55	0.94	3.51–3.59
Tangibles	3.61	1.06	3.56–3.65
Reliability	3.57	1.08	3.53–3.62
Responsiveness	3.38	1.26	3.33–3.44
Assurance	3.69	1.09	3.65–3.74
Empathy	3.47	1.25	3.42–3.52
Quality of health communication	2.83	1.17	2.78–2.88

**Fig 4.3 :**Descriptive statistics for continous variable.

## **PROS:**

- People acquire more information, both in terms of quality and quantity, as the quality of health communication improves; their assessments will be more informed and reasonable.
- Regardless of the country's economic progress, the family is the primary unit for health education in all countries.

## **CONS:**

When someone is less healthy (in this case, the danger of becoming overweight), they are more prone to project their own concerns onto the rest of society, leading to a negative view of community health.

## **CONCLUSION:**

Better access to health-related knowledge has been found to significantly affect a patient's view of health-care service quality, whether through mass-media health communication or first-hand or second-hand medical care experiences from family and friends.

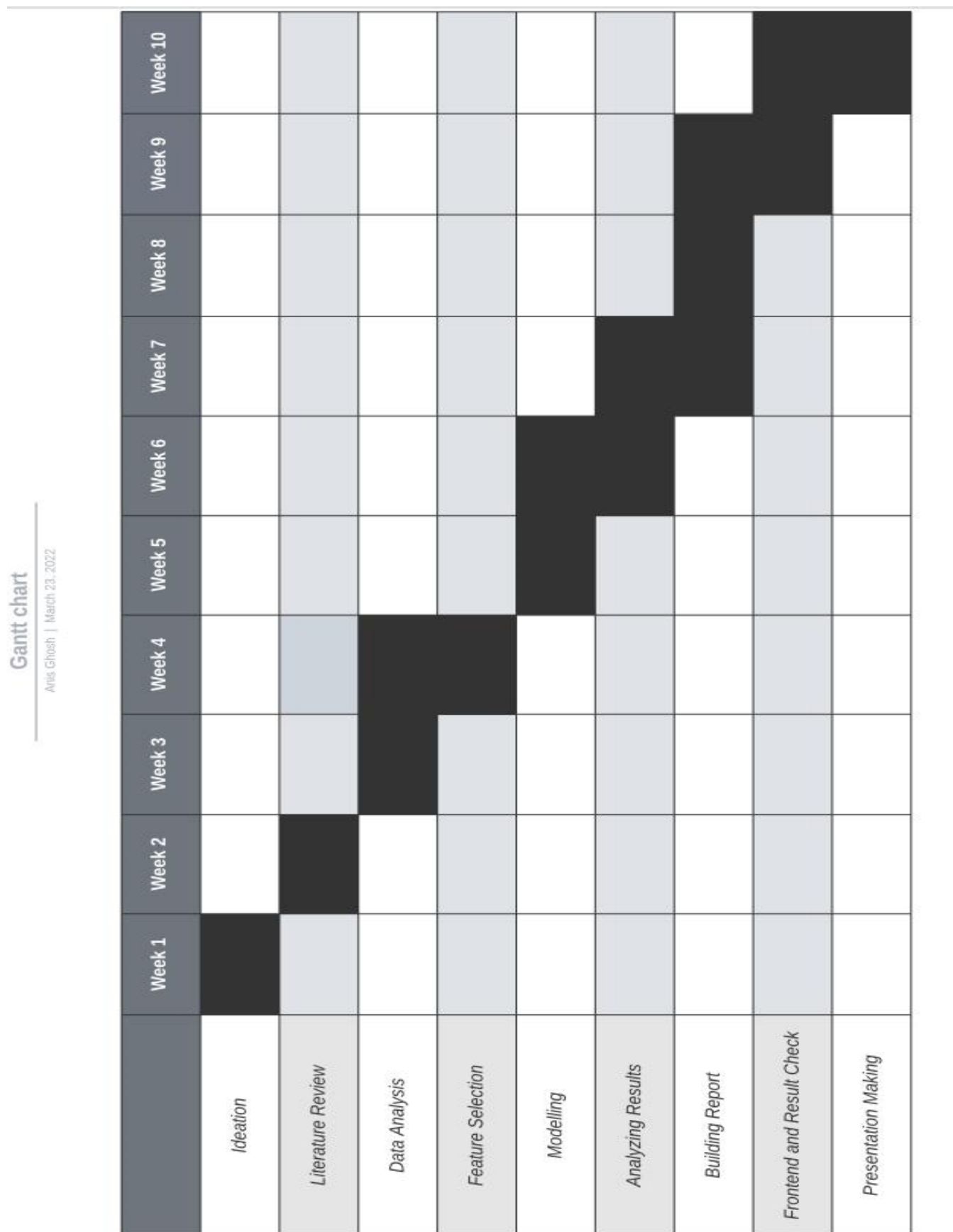
People in better physical shape had a more optimistic view of public health; but, contrary to common opinion, favourable evaluations of health-care service quality are linked to a pessimistic view of public health, and vice versa.



### 5.1. PLANNING AND PROJECT MANAGEMENT :

Activity	Starting week	Number of weeks
Ideation	1st week of January	1
Literature Review	2nd week of January	1
Data Analysis	3 <sup>rd</sup> week of January	2
Feature Selection	1 <sup>st</sup> week of February	1
Modelling	2 <sup>nd</sup> Week of February	2
Analyzing Results	4 <sup>th</sup> Week of February	2
Building Report	1 <sup>st</sup> Week of March	3
Frontend and Result Check	4 <sup>th</sup> Week of March	1
Report Modification	1 <sup>st</sup> Week of April	1
Preparation of project presentation	2 <sup>nd</sup> Week of April	1

The Gantt chart is shown below:



## References:

<https://data.worldbank.org/indicator/pa.nus.fcrf>

<https://developer.nytimes.com>

<https://developer.nytimes.com>

[GitHub - HS189/FinancialArbitrage2: sadbois](#)

<https://stackoverflow.com>

<https://www.youtube.com/watch?v=mrExsjcvF4o>

[7] GONGGI, S., 2011. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. In: Proceedings of the 3 rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

[8] LEXPRESS.MU ONLINE. 2014. Available from: <http://www.lexpress.mu/> [Accessed 17 January 2014]

[9] LE DEFI MEDIA GROUP. 2014. Available from: <http://www.defimedia.info/> [Accessed 17 January 2014]

[10] GELMAN, A. AND HILL, J., 2006. Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, New York, USA. 764 Sameerchand Pudaruth

[11] WEKA 4: DATA MINING SOFTWARE IN JAVA. 2014. Available from: <http://www.cs.waikato.ac.nz/ml/weka/index.html> [Accessed 17 January 2014].

[12] LI, Y. H. AND JAIN, A. K., 1998. Classification of Text Documents. The Computer Journal, Vol. 41, pp. 537-546.

[13] QUINLAN, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

[14] MITCHELL, T. M., 1997. Machine Learning. McGraw-Hill, Inc. New York, NY, USA.

[15] Strauss, Oliver Thomas, and Morgan Scott Hansen. "Advanced data science systems and methods useful for auction pricing optimization over network." U.S. Patent Application No. 15/213,941.

- [16] [Xinyuan Zhang , Zhiye Zhang and Changtong Qiu, "Model of Predicting the Price Range of Used Car", 2017](#)
- [17] [W.A. Awad and S.M. ELseuofi, "Machine Learning Method for SpamEmail Classification", 2011](#)
- [18] [Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification Method using Support Vector Machine", 2009](#)
- [19] [Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4.7 \(2014\): 753-764.](#)
- [20] [Noor, Kanwal, and Sadagat Jan. "Vehicle Price Prediction System using Machine Learning Techniques." International Journal of Computer Applications 167.9 \(2017\).](#)
- [21] [Kuiper, Shonda. "Introduction to Multiple Regression: How Much Is Your Car Worth?." Journal of Statistics Education 16.3 \(2008\)](#)
- [22] [http://ripublication.com/irph/ijict\\_spl/ijictv4n7spl\\_17.pdf](http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf)
- [23] [Gareth, J., Daniela, W., Trevor, H., & Tibshirani, R. \(2013\). An Introduction to Statistical](#)
- [24] [Learning \(Vol. 8\). https://doi.org/10.1016/j.peva.2007.06.006](#)
- [25] [Hurwitz, E., & Marwala, T. \(2012\). Common mistakes when applying computational intelligence and machine learning to stock market modelling. arXiv preprint arXiv:1208.4429.](#)
- [26] [Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. \(2018, April\). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. In Future of Information and Communication Conference \(pp. 413–422\). Springer, Cham.](#)
- [27] [Raschka, S., & Mirjalili, V. \(2017\). Python machine learning. Packt Publishing Ltd.](#)
- [28] [Agencija za statistiku BiH. \(n.d.\), retrieved from: http://www.bhas.ba . \[accessed July 18, 2018.\]](#)
- [29] [Listiani, M. \(2009\). Support vector regression analysis for price prediction in a car leasing application \(Doctoral dissertation, Master thesis, TU Hamburg-Harburg\).](#)
- [30] [Richardson, M. S. \(2009\). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /cocco%3A1346 \[accessed: August 1, 2018.\] \[4\] Wu,](#)

- J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications*, 36
- [31] 7809-7817. [5] Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. *Marketing Science*, 28(4), 637-644.
- [32] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on* (Vol. 2, pp. 682-685). IEEE.
- [33] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, 4(7), 753-764.
- [34] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*, 167
- [35] , 27-31. [9] Auto pijaca BiH. (n.d.), Retrieved from: <https://www.autopijaca.ba>. [accessed August 10, 2018].
- [36] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: <https://www.cs.waikato.ac.nz/ml/weka/>. [August 04, 2018].
- [36] [http://ripublication.com/irph/ijict\\_spl/ijictv4n7spl\\_17.pdf](http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf)
- [37] Gareth, J., Daniela, W., Trevor, H., & Tibshirani, R. (2013). An Introduction to Statistical
- [38] Learning (Vol. 8). <https://doi.org/10.1016/j.peva.2007.06.006>
- [39] Hurwitz, E., & Marwala, T. (2012). Common mistakes when applying computational intelligence and machine learning to stock market modelling. *arXiv preprint arXiv:1208.4429*.
- [40] Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. In *Future of Information and Communication Conference* (pp. 413–422). Springer, Cham.
- [41] Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Packt Publishing Ltd.
- [42] Agencija za statistiku BiH. (n.d.), retrieved from: <http://www.bhas.ba> . [accessed July 18, 2018.]

- [43] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).
- [44] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: <https://digitalcc.coloradocollege.edu/islandora/object/cocccc%3A1346> [accessed: August 1, 2018.] [4] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications*, 36 7809-7817.
- [45] Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. *Marketing Science*, 28(4), 637-644.
- [46] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on* (Vol. 2, pp. 682-685). IEEE.
- [47] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, 4(7), 753-764.
- [48] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*, 167 27-31.
- [49] Auto pijaca BiH. (n.d.), Retrieved from: <https://www.autopijaca.ba>. [accessed August 10, 2018].
- [50] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: <https://www.cs.waikato.ac.nz/ml/weka/>. [August 04, 2018]

