**Final Term Project**

**Introduction to Data Science**

**Section: A**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **ID** |
| **RABBI HOSSEN** | **20-44220-3** |
| **UDDIN, MD BORHAN** | **20-44002-2** |

**Submitted to: TOHEDUL ISLAM**

**Dataset Description:**

The Panic Disorder Detection Dataset comprises a total of 120,000 records focusing on individuals diagnosed with Panic Disorder. The dataset is divided into two key files, with the "Panic\_Disorder\_training" file intended for training machine learning models. This training file encompasses 100,000 labeled records, providing a substantial amount of data for model development.

**Sources of data :** Ayesha Clinic (Panjab)

**Collection Methodology :** Data is gathered from Documents Observations.

**Attributes:** The dataset includes a set of attributes that capture various aspects of individuals experiencing Panic Disorder. The key attributes are:

**Participant ID:** A unique identifier for each participant.  
**Age:** Participant's age at the time of data collection.  
**Gender:** Participant's gender identity.  
**Family History:** Whether the participant has a family history of panic disorder or related mental health conditions.  
**Personal History:** Information about the participant's past diagnoses, mental health experiences, and potentially significant life events.  
**Current Stressors:** Current sources of stress in the participant's life.  
**Symptoms:** Specific panic disorder symptoms reported by the participant.  
**Severity:** The severity of the participant's panic disorder symptoms, potentially represented by a scoring system.  
**Impact on Life:** The extent to which the participant's panic disorder impacts their daily life and functioning.  
**Demographics:** Additional demographic information about the participant, such as ethnicity, socioeconomic status, etc. (not explicitly confirmed in the provided information).

**Medical History:** Existing medical conditions and treatments.  
**Psychiatric History:** Previous diagnoses of mental health disorders.  
**Substance Use:** Information about tobacco, alcohol, or other substance use.  
**Coping Mechanisms:** Strategies used to manage stress and symptoms.  
**Social Support:** Availability and quality of social support networks.  
**Lifestyle Factors:** Diet, exercise, sleep habits, etc.

**Panic Disorder Diagnosis:** Confirmed diagnosis of panic disorder through clinical evaluation

 Dataset Link : [Panic Disorder Detection Dataset (kaggle.com)](https://www.kaggle.com/datasets/muhammadshahidazeem/panic-disorder-detection-dataset/)

**Convert XLSX to CSV and Import CSV and print the dataset:**

**A screenshot of a computer

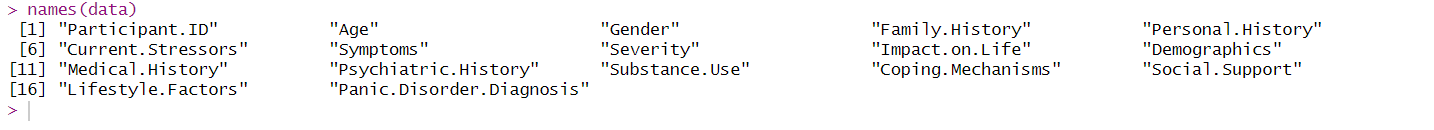
Description automatically generated**

**Description:**

The code is for read the dataset and also to prove the successful conversion of a xlsx file to csv file. Also, the output of the dataset by using the print()

**Column Name showing :**

names(data)

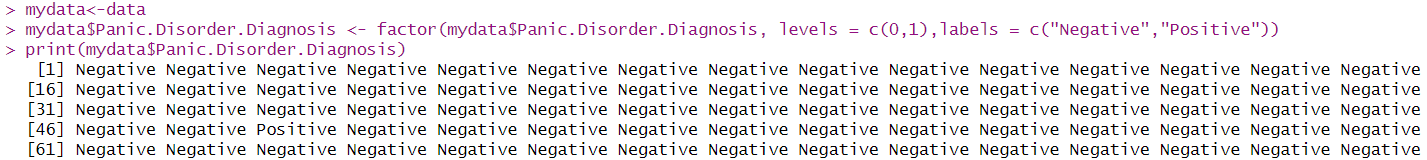
****

**Description:**

Use ‘Names()’ function to see columns name of the dataset.

**Dataset preprocessing:**

**Conversion (numerical to categorical):**

****

**A close-up of a computer screen

Description automatically generated**

**Description:**

We convert categorical value to numerical value for getting accurate mathematic calculation. Levels and labels are used for transform characters to integer and bind them and “factor()” function  is used to encode a vector as a factor. Use ifelse to categories a range of number.

**Missing value checking:**

A screenshot of a computer

Description automatically generated

**Description:**

useNA = "ifany" is used for checking if there any missing value in the categorical dataset. If there any missing value, then it will show it in a new column “NA”

**Correlation:**

A screenshot of a computer code

Description automatically generated A screenshot of a computer code

Description automatically generated

**A screenshot of a computer

Description automatically generated** **A screenshot of a computer

Description automatically generated**

A computer screen shot of a computer code

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated A computer code with text

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer program

Description automatically generated

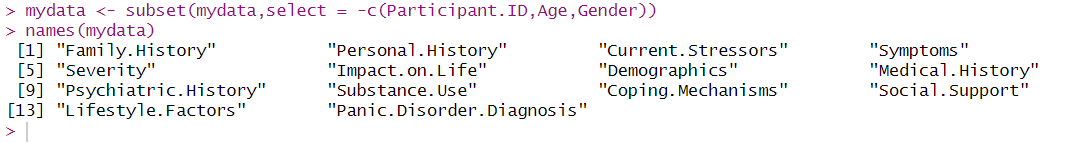
A screenshot of a computer code

Description automatically generated

**Description:**

The function data.frame() creates data frames, tightly coupled collections of variables which share many of the properties of matrices and of lists, used as the fundamental data structure. “=” is used for Assigning a value to a name. “table()” forming a table. “chisq.test” performs chi-squared contingency table tests and goodness-of-fit tests.

**creating Significant attributes in dataset for naïve baye classification :**



**Description:**

The “subset()”  function in R is used to extract a subset of a data frame based on specified conditions. “select =-c()”  is used for selecting columns without -c() attributes.

In correlation, we know that significant value is 0.05 or less . In this dataset we find “participant.Id, Age, Gender” have greater correlation than 0.05 .so we discard this three attributes from the dataset.

**Naïve baye classification method -1 (Dividing the data into training and test set):**

**Install pakages & laod library:**

**A black text on a white background

Description automatically generated**

**Description:**

Download and install packages from CRAN-like repositories or from local files. “library” is used to load .

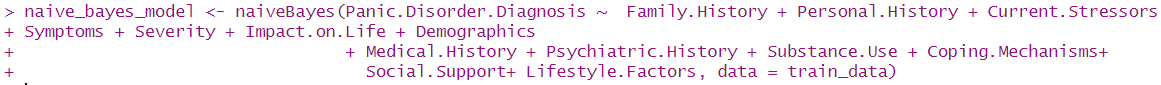
**Split the dataset into train and testing set :**

A close-up of a code

Description automatically generated

**Description:**

The “set.seed” function in R is used to set the seed for the random number generator. This ensures reproducibility in statistical analyses that involve random processes, such as random sampling or generating random numbers. this code is performing a random split of the “your\_data” dataset into a training set “(train\_data)” and a test set “(test\_data)”. The training set contains 70% of the original data, randomly sampled, while the test set contains the remaining 30%. This is a common approach for creating training and test sets for machine learning models to evaluate their performance on unseen data.

****

**Description:**

“naiveBayes()” Computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule.

“~”  Tilde is used to separate the left- and right-hand sides in a model formula. Here left side is target attribute and right side is feature attribute. For adding more feature attribute we use “+” .

****

**Description:**

“predict” is a generic function for predictions from the results of various model fitting functions. The model uses the learned patterns from the training data to make predictions on the new, unseen data in the test\_data dataset.

**A close-up of a computer code

Description automatically generated**

**Description:**

“diag()” is used for creating diagonal matrix and “sum()” is used for summation of the matrix.

Finally calculate accuracy of the classification. The accuracy is 0.963 ~ 96% .

**Naïve baye classification method -2 ( 10-fold cross validation) :**

**Install pakages & laod library:**

A close up of a text

Description automatically generated

**Description:**

We use “boxplot()” function to find out outlier in graph. “boxplot.stats()$out” function is used to know which number are outlier.

A close-up of a computer code

Description automatically generated

**Description:**

It specifies the method as "cv" for cross-validation and the number of folds as 10. “~ “ indicates that Panic.Disorder.Diagnosis is the target variable, and “.” includes all other variables in the dataset as predictors. “naive\_bayes\_model\_cv$results” Accesses the results of the cross-validation, including various performance metrics for each fold. “cv\_accuracy <- mean(cv\_results$Accuracy):” Calculates the average accuracy across all folds using the accuracy values from “cv\_results”.

Finally we find the accuracy is 0.8312~ 83%

A screenshot of a computer code

Description automatically generated

**Description:**

we find the best accuracy for classification is 96% .We extracts values from the confusion matrix to compute recall, precision, and F-measure for classification problem where the positive class is represented by index 2 in the matrix.