**3. METHODOLOGY**

* 1. **DATA**: In this paper data, which is used, is primary data. This means the data is collected from the authors' end. It’s a big challenge to find the right people and collect all the relevant information about their mental health like facing Anxiety, Depression, Eating disorders, hallucinations, Excessive mental stress for studies, etc. Nearly 210 samples and 97 attributes are collected from the author's end during this work from the local area of west Bengal.
  2. **DATA CLEANING**: Though it's a primary dataset, many difficulties must be faced, like some fields are blank and some fields are filled by garbage value. So to successfully perform the research works first step is to clean the data. Here authors filled the field by putting the mean value of this attribute’s value.
  3. **Finds Most Relevant Attributes**: After cleaning the data, the processing part of the dataset is completed. The next part is to identify the correct attributes which are the most relevant and mostly dependent for having **‘thoughts of self-harm or suicide’**. Here author finds that from the 97 attributes( 1 is Predicting Attribute) only 24 attributes are the most relevant attribute, which directly impacts ‘Thought of self-harm or suicide’.
  4. **Table 1: The attributes along with datatypes**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **Datatypes** | **Mean** | **Description** |
| 'Your gender[ Male=1 , Female=0 ]' | Integer (0/1) | 0.41 | Male=1  Female=0 |
| 'College student' | Integer(0/1) | 0.70 | College student=1  Others=0 |
| 'Anxiety' | Integer(0/1) | 0.62 | Has Anxiety=1  Hasn’t Anxiety=0 |
| 'Depression' | Integer(0/1) | 0.62 | Has Depression=1  Hasn’t Depression=0 |
| 'Separated' | Integer(0/1) | 0.01 | Separated=1  Not Separated=0 |
| 'Eating disorders( Anorexia, Bulimia, Binge eating etc.)' | Integer(0/1) | 0.21 | Has Eating disorders=1  Hasn’t Eating disorders=0 |
| 'Faced this situation currently' | Integer(0/1) | 0.47 | Faced this situation currently=1  Faced this situation in past=0 |
| 'Facing these problems less than 1 month' | Integer(0/1) | 0.13 | Facing these problems less than 1 month—  Yes=1  No=0 |
| 'Hallucination' | Integer(0/1) | 0.04 | Has Hallucination=1  Hasn’t Hallucination=0 |
| 'Fatigue' | Integer(0/1) | 0.29 | Fatigue=1  Not Fatigue=0 |
| 'Guilt without any reasons' | Integer(0/1) | 0.41 | Guilt without any reasons----  Yes=1  No=0 |
| 'Drug and alcohol misuse' | Integer(0/1) | 0.03 | Drug and alcohol misuse----  Yes=1  No=0 |
| 'Excessive mental stress for studies' | Integer(0/1) | 0.34 | Excessive mental stress for studies-----  Yes=1  No=0 |
| 'Social disadvantage, poverty or debt' | Integer(0/1) | 0.08 | Social disadvantage, poverty or debt—  Yes=1  No=0 |
| 'Being a long-term carer for someone] | Integer(0/1) | 0.07 | Being a long-term carer for someone-----  Yes=1  No=0 |
| 'Being involved in a serious incident in which you feared for your life’ | Integer(0/1) | 0.12 | Being involved in a serious incident in which you feared for your life------  Yes=1  No=0 |
| 'Is your problem diagnosed ?' | Integer(0/1) | 0.91 | Is your problem diagnosed ?---  Yes=1  No=0 |
| 'Are you still in treatment ?' | Integer(0/1) | 0.13 | Are you still in treatment ?----  Yes=1  No=0 |
| 'More than 1 month and less than 6 months in treatment ' | Integer(0/1) | 0.08 | More than 1 month and less than 6 months in treatment----  Yes=1  No=0 |
| 'Is your problem fixed with help from friends or relatives’ | Integer (0/1) | 0.16 | Is your problem fixed with help from friends or relatives-----  Yes=1  No=0 |
| 'Is your problem fixed by self-realization or self-motivation ’ | Integer(0/1) | 0.38 | Is your problem fixed by self-realization or self-motivation---------  Yes=1  No=0 |
| 'Trust Issues | Integer(0/1) | 0.23 | Trust Issues-----  Yes=1  No=0 |
| 'Health Crisis' | Integer(0/1) | 0.07 | Health Crisis-----  Yes=1  No=0 |
| 'Thought of self-harm or suicide’ | Integer(0/1) | 0.29 | Thought of self-harm or suicide-----  Yes=1  No=0 |

**4. RESEARCH METHOD**

**4.1 Multiple linear regression :**

Multiple linear regression is the method of statistics in regression that is used to analyze the relationship between a single response variable (dependent variable) with two or more controlled variables (independent variables). This method was selected for this research because there was more than one controlled variable. In this research, the response variable is **‘Thought of self-harm or suicide (Y)’** while the state

(X1): 'Your gender[ Male=1 , Female=0 ]',(X2): 'College student', (X3): ‘Anxiety’, (X4): ‘Depression’, (X5): ‘Separated’, (X6): ‘Eating disorders’, (X7): ‘Faced this situation currently’, (X8): ‘Facing these problems less than 1 month’, (X9): ‘Hallucination’, (X10): ‘Fatigue’, (X11): ‘Guilt without any reasons’, (X12): ‘Drug and alcohol misuse’, (X13): ‘Excessive mental stress for studies’, (X14): ‘Social disadvantage, poverty or debt’, (X15): ‘Being a long-term carer for someone’, (X16): ‘Being involved in a serious incident in which you feared for your life’ (X17): ‘Is your problem diagnosed ?’, (X18): ‘Are you still in treatment ?’, (X19): ‘More than 1 month and less than 6 months in treatment’, (X20): ‘Is your problem fixed with helped by friends or relatives’, (X21): ‘Is your problem fixed by self realization or self motivation’ (X22): ‘Trust Issues’, and (X23): ‘Health Crisis’ are controlled variables.

**Multiple linear regression formula:**

**Ylinear = a+𝑏1𝑋1 + 𝑏2𝑋2 + … … … …. +𝑏𝑛𝑋𝑛**

Where,

bi = ∑( xi – x)(yi – y) / ∑( xi – x )2 [Coefficient of the control variable]

a = yi -( **𝑏1𝑋1 + 𝑏2𝑋2 + … … … …. +𝑏𝑛𝑋𝑛 ) [**Constant variable]

y= Dependent variable

xi = controlled variables(independent variables)

x = controlled variables mean value

y = Dependent variable mean value

**4.2 Logistic regression :**

Logistic Regression is a statistical method used for binary classification problems, where the goal is to predict one of two possible outcomes. Unlike linear regression, which is used for continuous dependent variables, logistic regression is used when the dependent variable is dichotomous, meaning it has two possible values, such as "yes" or "no", "success" or "failure", or "dead" or "alive". In this research, 0<= ylinear<=1 so the authors decided to use logistic regression here.

The formula for Logistic Regression :

ylogistic =1/(1+e-ylinear)

where,

e=2.718

ylinear = Multiple Linear Regression value

**4.3 Confusion-Matrix**

After finding the accuracy of the difference between actual data and calculated data we did the Confusion Matrix. In this confusion matrix it can be seen that we find the **TP** – which stands for ‘**TRUE POSITIVE’** means the accuracy of classified positive data, **TN** – which stands for ‘**TRUE NEGATIVE’** means the accuracy of classified negative data, **FP** – which stands for ‘**FALSE POSITIVE’**, means which remark that actual value is negative but predicted data is positive, **FN**– which stands for ‘**FALSE NEGATIVE’** means that actual data and the predicted data both are negative and append the TP, TN, FP, FN value in 2\*2 matrix(mat1). After that, we find the accuracy, sensitivity, precision, recall, and specificity. This matrix contains all the raw information about the predictions done by a classification model on a given data set.

**4.4Cross-Validation**

After finding the accuracy of the difference between actual data and calculated data we did cross-validation. In this cross-validation process first, we divide the whole list into 10 sub-list and then we find the accuracy of 10 sub-list elements we also find the Confusion Matrix of each Sub-list and we find the accuracy, and sensitivity, precision, recall, and specificity.

ACCURACY: It’s the ratio of the correctly labeled subjects to the whole pool of subjects. Accuracy is intuitional.

PRECISION: Precision is the ratio of the correctly +velabeled by our program to all +velabeled.

SENSITIVITY: Sensitivity means out of the total positive, what percentage are predicted positive.

SPECIFICITY: Specificity is calculated as the number of correct negative predictions divided by the total number of negatives.

* **ACCURACY= (TP+TN/ TP+TN+FP+FN)\* 100**
* **PRECISION = (TP/FP+TP)\*100**
* **SENSITIVITY= (TP/FN+TP)\*100**
* **SPECIFICITY = (TN/TN+𝐹𝑃)\* 100**

**F1 score:** - The F1 score is defined as the harmonic mean of precision and recall. As a short reminder, the harmonic mean is an alternative metric for the more common arithmetic mean. It is often useful when computing an average rate. In the F1 score, we compute the average precision and recall. For, finding the F1 score we have to find the first precession and recall

**Precession = TP / (TP+FP)**

**Recall = TP /(TP+FN)**

**F1 Score = (2∗𝑅𝑒𝑐𝑎𝑙𝑙∗𝑝𝑟𝑒𝑐𝑒𝑠𝑠𝑖𝑜𝑛) / (𝑅𝑒𝑐𝑎𝑙𝑙+𝑃𝑟𝑒𝑐𝑒𝑠𝑠𝑖𝑜𝑛)**

**5. RESULTS**

5.1) **Confusion-Matrix: -**

**Table-2**

**Taking 80 Percent Data as Training and 20 Percent as Test:**

|  |  |
| --- | --- |
| Confusion Matrix | 6 2  5 29 |
| Accuracy | 83.33 |
| Sensitivity | 75.0% |
| Specificity | 85.29% |
| Precision | 0.54 |
| Recall | 0.75 |
| F1 Score | 0.63 |
| kappa | 1 |

**Table-3**

**Taking 66 Percent Data as Training and 34 Percent as Test:**

|  |  |
| --- | --- |
| Confusion Matrix | 10 2  5 29 |
| Accuracy | 86.11 |
| Sensitivity | 75% |
| Specificity | 85.29% |
| Precision | 0.54 |
| Recall | 0.75 |
| F1 Score | 0.63 |
| kappa | 1 |

**Table-4**

**Taking 50 Percent Data as Training and 50 Percent as Test**

|  |  |
| --- | --- |
| Confusion Matrix | 26 2  17 60 |
| Accuracy | 81.90 |
| Sensitivity | 83.33% |
| Specificity | 86.66% |
| Precision | 0.55 |
| Recall | 0.83 |
| F1 Score | 0.66 |
| kappa | 1 |

5.2) **For 10-fold cross-validation:**

**Table-5**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test Cases | Accuracy Rate | Sensitivity | Specificity | Precession | Recall | F1 score | Kappa |
| 01 | 90.47 | 66.44% | 94.44% | 0.66 | 0.66 | 0.66 | 0.61 |
| 02 | 90.47 | 0.0% | 95.0% | 0.0 | 0.0 | 0.0 | -0.04 |
| 03 | 80.95 | 0.0% | 94.44% | 0.0 | 0.0 | 0.0 | -0.07 |
| 04 | 71.42 | 76.92% | 62.5% | 0.76 | 0.76 | 0.76 | 0.39 |
| 05 | 76.19 | 83.33% | 66.66% | 0.76 | 0.83 | 0.8 | 0.50 |
| 06 | 76.19 | 85.71% | 57.14% | 0.8 | 0.85 | 0.82 | 0.44 |
| 07 | 85.71 | 50.0 % | 100.0% | 1.0 | 0.5 | 0.66 | 0.58 |
| 08 | 85.71 | 100.0 % | 85.0 % | 0.35 | 1.0 | 0.4 | 0.35 |
| 09 | 90.47 | 100.0% | 90.47% | 0.33 | 1.0 | 0.5 | 0.46 |
| 10 | 76.19 | 66.66% | 80.00% | 0.57 | 0.66 | 0.61 | 0.44 |