International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

Prediction of suicidal tendencies using machine learning

Gourab Karak[1], Kishalay Ghosh[2], Nur Muktada Hasan Ansari[3], Mrs. Sulekha Das[4], Dr. Avijit Kumar Chaudhuri[5]
[1]UG-Computer Science and Engineering, Techno Engineering College Banipur
[2]UG-Computer Science and Engineering, Techno Engineering College Banipur
[3]UG-Computer Science and Engineering, Techno Engineering College Banipur
[4]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur
[5]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur
Corresponding Author Orcid ID: [1]0000-0002-5173-7163, [2]0000-0002-5312-8820, [3]0000-0001-7935-3199, [4]0000-0002-6641-3268, [5]0000-0002-5310-3180

ABSTRACT

Suicide is nowadays a crucial public health issue. In various communities, people are much more affected by this severe tendency. Nowadays, suicide connects to mental illness. Along with depression, drug misuse is an equally important risk factor. Although anxiety, personality, and trauma-related issues, as well as genetic and mental disorders, also equally contributed to this health issue. This research work conducts to find the relation among suicidal tendencies, depression, anxiety, daily stresses, and mental health among all people, including employees, non-employee, students, old age, and all others. Step-by-step multiple linear regression, logistic regression analyses, confusion matrix, and cross-validation use for the data analysis. Suicidal illusions have a significant and positive relation with depression, anxiety, mental stress, and drug use disorder. In this research work regression technique showed that depression is the most crucial feature in the prediction of suicidal behavior, where anxiety, mental health, and daily stresses are the next, followed by respectively. Psychological problems and mental health issues with other factors such as gender problems, marital problems, and hallucinations play a role in suicidal thoughts [1]. Through this research, approximately 84 percent (%) of the data predict correctly.

Keywords— Mental Health, Suicidal Tendencies, Multiple Linear Regression, Logistic Regression, Confusion Matrix, Cross-Validation

1. Introduction

Suicide is known as a fundamental general health problem in the world and is a disturbance and pathetic issue for psychotherapists, psychologists, and other specialists, involved in people's welfare and psychological health problem. Suicide is also considered the third major cause of death among adolescents and youths [1]. On the whole, older adults maintain emotional well-being compared to those at younger ages. Yet, at the same time, adults aged 70 and older have the highest rates of suicide deaths in most regions of the world (World Health Organization, 2014). As another example, there appear to be not just age effects on suicide risk [5].

Machine learning can simply be defined as using data instead of logic to perform tasks by a machine. We use the data to train the machine, as in, tell it what it has to do and then test the trained model on different tasks to see whether the training has been successful or not. When it comes to data mining, the term classification plays an important role as it assigns class values to new instances found during data mining.

Public consultation of risk for people's suicidal exploits is often constituted by the primacy of contributions from depression. Now teenagers are much more affected by suicidal trauma. For example, the New Zealand Youth Suicide Prevention Strategy published by the Ministries of Youth Affairs, Health, and Maori Affairs for dissemination to the general population comments: "Around 90% of young people dying by suicide or making suicide attempts are likely to have one or more recognizable psychiatric disorders at the time" [2].

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

While acknowledging that the risk factors contributing to suicidal behaviors among young people are complex and multifactorial, a review of the suicide literature suggests a relatively uncomplicated relationship between young people's suicidal behaviors and depression. Additionally, recent research suggests that depression is likely to be significantly under-diagnosed in men, such that men experiencing major depression are much more likely to suicide because they are consistently less able to access help than their female counterparts.

The risk factors of this problem are psychological disorders and socio-biological factors. 1.5 of all deaths in both males and females occur because of self-simulated injuries and is in the second order of the two major causes of death in the 15-34 age level and the 10th cause of death in all age levels all over the world. There is some different factors which increase the mentality of attempting suicide among the youths which range from the lowest risk level factors to the highest ones and are put in categories such as personal, family, demographic, social environment, and daily stress factors. Low-risk factors derive from a small-scale affair in family and academic condition and end in now-and-then feelings of regret without the individual's having depression previously. Those who are in the middle ground of risk for endeavoring suicide have some suicidal thoughts, habitual depression, anxiety, and temper control issues. Some have the mentality of attempting suicide in highly manner who may intoxicate from their homes, have no meaning of continuing their life and think, it will be better if they have died. Other factors which draw up the ground for attempting suicide are high tension and intense stress. Research came to end that 94.8% of university students who had thoughts of suicide had been so depressed in their educational year that they felt no inspiration nor function and 94.4% of them believed that hopelessness punctured into the depth of their existence. The personality attributes that have an association with suicidal thoughts in one way or another are low levels of self-esteem, outside control, helplessness, impulsion, aggression, and sociability. Above 90% of those who were successful doing suicide had at least one mental disorder at the time of death such as temper loss, anxiety disorders, and anti-social behaviors [1].

So it is considered necessary for those who are in-charge in the institutions to teach the students' resiliency skills and to diagnose and solve their mental problems. To promote mental health level and prevent their thinking about suicide and attempting it.

2. Literature Review

Logistic regression is an applied statistical technique. Target of logistic regression analysis is to use the independent variables whose worth square measure acknowledged to predict the worth of the only dependent value. The main aim of this research is the prediction the dependence of suicidal tendencies in various communities.

Mr. Mark A.ilgen had done (2008) on "Pain and suicidal thoughts, plans and attempts in the United States".

Adamson S Muula had done research (2007) on "Suicidal ideation and associated factors among school-going adolescents in rural Uganda".

Kara Zivin had done research (2009) on "Persistence of mental health problems and needs in a college student population".

R.D. Goodwin had the research (2002) based on "Major Depression, Physical Illness, and Suicidal Ideation in Primary Care".

Those and many others had done their research prediction using logistic regression.

We have gathered some specific ideas about machine learning and logistic regression. So we were very interested in doing a project based on it. So we collected some real-life data on the "Communities and Suicidal Data Set" and try to predict suicidal tendencies using independent fields like anxiety, depression, drug and alcohol misuse, etc.

3. Methodology

3.1 Data:: In this paper, the authors have used the primary dataset. 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

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

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.

3.2 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.

3.3 Find 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'**.

Table 1: Attributes along with data types

Attributes	Data Types	Mean	Description
'Your Gender [Male = 1, Female = 0]'	Binary (0/1)	0.41	Male = 1 Female = 0
'College Student'	Binary (0/1)	0.70	College Student =1 Others = 0
'Anxiety'	Binary (0/1)	0.62	Has anxiety = 1 Hasn't anxiety = 0
'Depression'	Binary (0/1)	0.62	Has depression = 1 Hasn't depression = 0
'Separated'	Binary (0/1)	0.01	Separated = 1 Not separated = 0
'Eating disorders (Anorexia, Bulimia, Binge eating, etc.)'	Binary (0/1)	0.21	Has eating disorders = 1 Hasn't eating disorders = 0
'Faced this situation currently'	Binary (0/1)	0.47	Faced this situation currently = 1 Faced this situation in the past = 0
'Facing these problems less than 1 month'	Binary (0/1)	0.13	Facing these problems less than 1 month --- Yes = 1 No = 0
'Hallucination'	Binary (0/1)	0.04	Has hallucination = 1 Hasn't hallucination = 0
'Fatigue'	Binary (0/1)	0.29	Fatigue = 1 Not fatigue = 0
'Guilt without any reasons'	Binary (0/1)	0.41	Guilt without any reason --- Yes = 1 No = 0
'Drug and alcohol misuse'	Binary (0/1)	0.03	Drug and alcohol misuse ---

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

			Yes = 1 No = 0
'Excessive mental stress for studies'	Binary (0/1)	0.34	Excessive mental stress for studies --- Yes = 1 No = 0
'Social disadvantage, poverty or debt'	Binary (0/1)	0.08	Social disadvantage, poverty, or debt --- Yes = 1 No = 0
'Being a long-term carer for someone'	Binary (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'	Binary (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?'	Binary (0/1)	0.91	Is your problem diagnosed? --- Yes = 1 No = 0
'Thought of self-harm or suicide'(**dependent variable**)	Binary (0/1)	**0.29**	Thought of self-harm or suicide --- Yes = 1 No = 0

4. Research Methodology

4.1 Multiple Linear Regression:: Multiple linear regression is the statistical method 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 [11].

Multiple Linear Regression Formula::

$$Y_{linear} = a + b_1X_1 + b_2X_2 + \dots \dots \dots \dots + b_nX_n$$

Where,

$b_i = \sum(x_i - \overline{x})(y_i - \overline{y}) / \sum(x_i - \overline{x})^2$ [coefficient of the control variable]

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

a = y_i - ($b_1X_1 + b_2X_2 + \dots\dots\dots\dots + b_nX_n$) [constant variable]
y = dependent variable
x_i = controlled variables (independent variables)
$\overline{x}$ = controlled variables mean value
$\overline{y}$ = dependent variable mean value

4.2 Logistic Regression:: Logistic Regression is a numerical method used for binary classification problems, where the target is to analyze 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, which means that it has two possible values, such as "yes" or "no", "success" or "failure", or "dead" or "alive". In this research, **0 <= Y_{linear} <= 1**, so we have decided to use logistic regression here.

The formula for Logistic Regression::

$$Y_{logistic} = 1 / (1 + e^{-y_{linear}})$$

Where,

e = 2.718
y_{linear} = 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 analyzed data both are negative and include the TP, TN, FP, FN value in **2*2 matrix** (mat1). After that, we find the accuracy, sensitivity, precision, recall, and specificity. This matrix holds all the raw data about the predictions done by a classification model on a collected data set.

4.4 Cross-Validation:: After checking the accuracy difference between the actual data and calculated data, the author is analyzing 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. After that, we find the confusion matrix of each sub-list and last we check the accuracy, sensitivity, precision, recall, and specificity [11].

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 + labeled by our program to all + labeled.

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:: It is described as the **harmonic mean of precision and recall**, where the harmonic mean is a substitute metric for the more common arithmetic mean. It is frequently useful when calculating a mean rate. In this phase, we find out 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)**

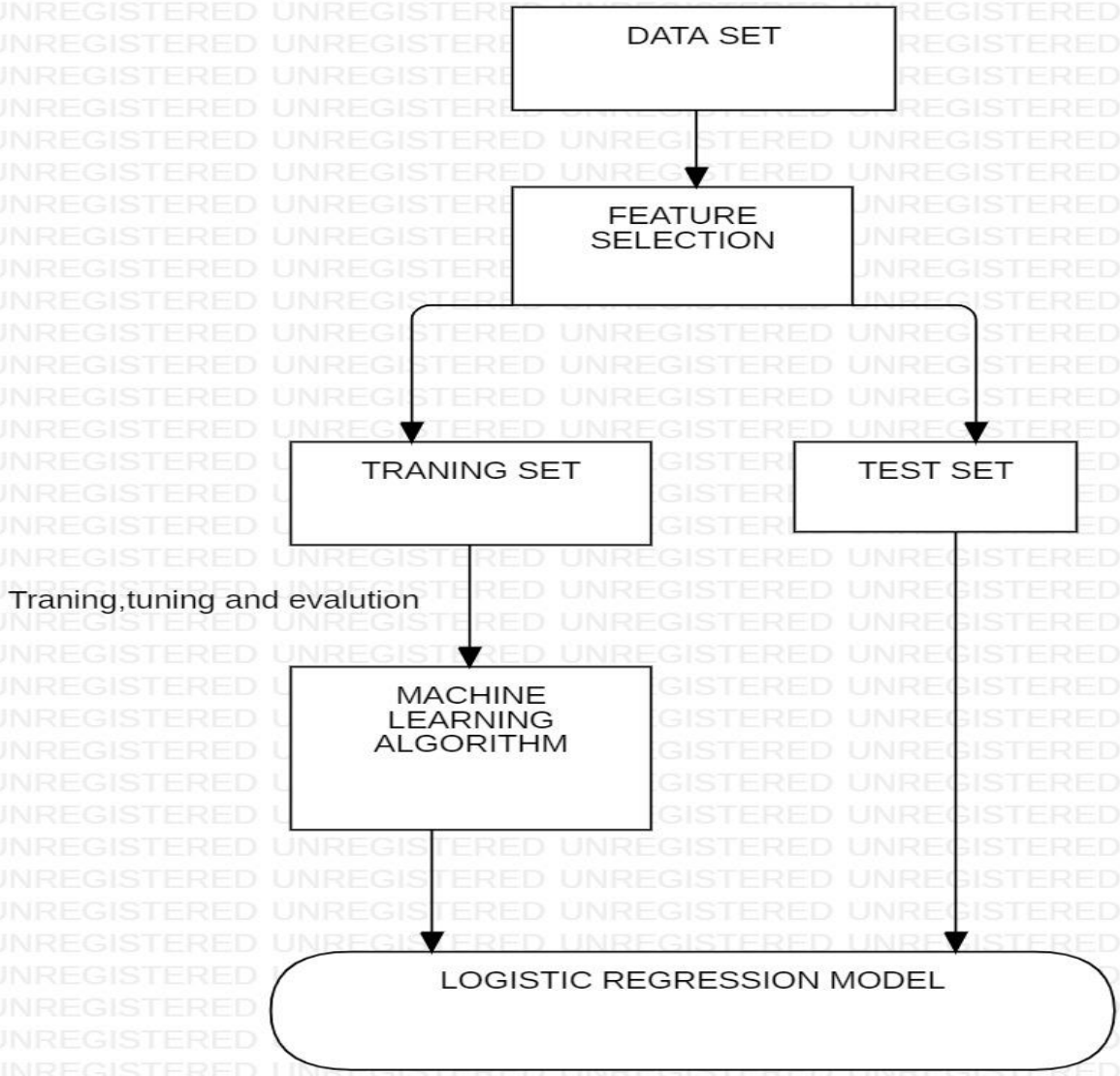
Kappa:: The kappa score is an interesting metric. Kappa or **Cohen's** Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. Cohen

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.-- $\mathbf{K = P_0 - P_e / 1 - P_e}$
where, P_0 = Probability of agreement, P_e = Probability of random agreement

4.5 Flow Chart::

Fig 1:



5. Results

5.1 Confusion Matrix::

Table 2: Taking 80% of Data as Training and 20% as a 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

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

Kappa	1

Table 3: Taking 66% of Data as Training and 34% as a 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 % of Data as Training and 50% as a 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

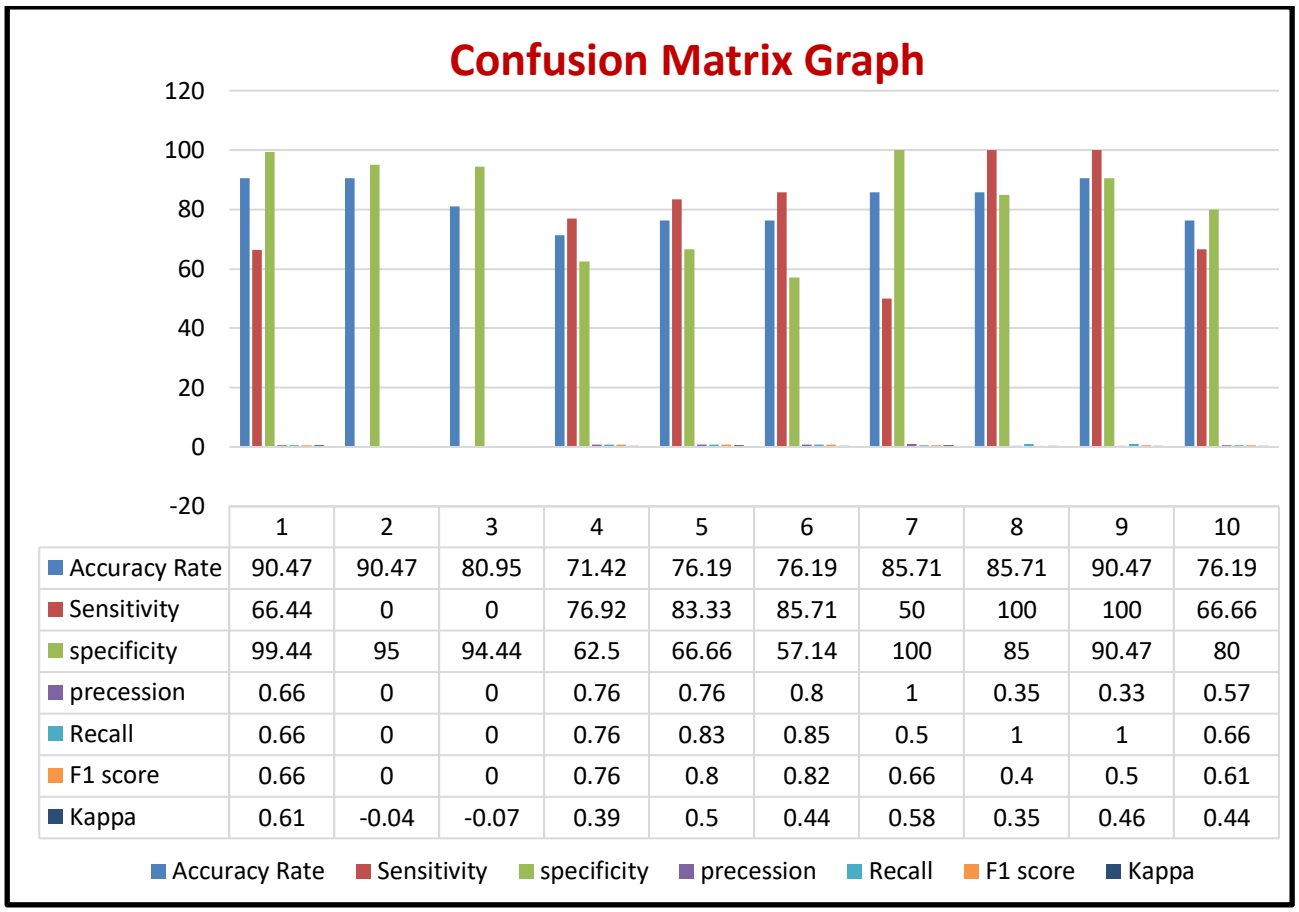
5.2 For 10-fold Cross-Validation::

Table 5:

Test Cases (Data Range)	Accuracy Rate	Sensitivity	Specificity	Precision	Recall	F1 score	Kappa
01 **(0 - 21)**	90.47	66.44%	94.44%	0.66	0.66	0.66	0.61
02 **(22 - 42)**	90.47	0.0%	95.0%	0.0	0.0	0.0	-0.04
03 **(43 - 64)**	80.95	0.0%	94.44%	0.0	0.0	0.0	-0.07
04 **(65 - 85)**	71.42	76.92%	62.5%	0.76	0.76	0.76	0.39
05 **(86 - 106)**	76.19	83.33%	66.66%	0.76	0.83	0.8	0.50
06 **(107 - 127)**	76.19	85.71%	57.14%	0.8	0.85	0.82	0.44
07 **(128 - 148)**	85.71	50.0 %	100.0%	1.0	0.5	0.66	0.58
08 **(149 - 169)**	85.71	100.0 %	85.0 %	0.35	1.0	0.4	0.35
09 **(170 - 190)**	90.47	100.0%	90.47%	0.33	1.0	0.5	0.46
10 **(171 - 210)**	76.19	66.66%	80.00%	0.57	0.66	0.61	0.44

Fig 2: 10-fold Cross-Validation Confusion Matrix Graph

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621



	1	2	3	4	5	6	7	8	9	10
Accuracy Rate	90.47	90.47	80.95	71.42	76.19	76.19	85.71	85.71	90.47	76.19
Sensitivity	66.44	0	0	76.92	83.33	85.71	50	100	100	66.66
specificity	99.44	95	94.44	62.5	66.66	57.14	100	85	90.47	80
precession	0.66	0	0	0.76	0.76	0.8	1	0.35	0.33	0.57
Recall	0.66	0	0	0.76	0.83	0.85	0.5	1	1	0.66
F1 score	0.66	0	0	0.76	0.8	0.82	0.66	0.4	0.5	0.61
Kappa	0.61	-0.04	-0.07	0.39	0.5	0.44	0.58	0.35	0.46	0.44

6. Conclusion

This paper is using logistic regression (LR) to predict suicide tendencies. We have collected the data from real life based on that we made a relationship between the dependent variable and the independent variable. After that, we perform cross-validation for more accuracy. After checking the cross-validation, we move to the confusion matrix where we compare the actual target values with those predicted by the machine learning model. Using these models, we predict the accuracy as well as sensitivity and specificity for 2/3rd, 1/2nd, and 4/5th sets of data. In this project, we define some actual reasons for increasing suicide. We have done this project because day-by-day this kind of suicide rate is increasing in our society. It will help any country's govt. to analyze their country's suicidal rate along with their mental health and they can also reduce the reasons behind increasing suicidal issues.

7. References

1. Izadinia, N., Amiri, M., ghorban Jahromi, R., & Hamidi, S. (2010). A study of the relationship between suicidal ideas, depression, anxiety, resiliency, daily stresses, and mental health among Tehran university students. *Procedia-Social and Behavioral Sciences*, *5*, 1615-1619.

2. Bennett, S., Coggan, C., & Adams, P. (2003). Problematizing depression: young people, mental health and suicidal behaviors. *Social Science & Medicine*, *57*(2), 289-299.

3. Lee, K. H., Jun, J. S., Kim, Y. J., Roh, S., Moon, S. S., Bukonda, N., & Hines, L. (2017). Mental health, substance abuse, and suicide among homeless adults. *Journal of evidence-informed social work*, *14*(4), 229-242.

4. Betz, M. E., Valley, M. A., Lowenstein, S. R., Hedegaard, H., Thomas, D., Stallones, L., & Honigman, B. (2011). Elevated suicide rates at high altitudes: sociodemographic and health issues may be to blame. *Suicide and Life-Threatening Behavior*, *41*(5), 562-573.

DOI:10.46647/ijetms.2023.v07i02.006 ISSN: 2581-4621

5. Van Orden, K. A., & Conwell, Y. (2016). Issues in research on aging and suicide. *Aging & mental health*, *20*(2), 240-251.
6. Brådvik, L. (2018). Suicide risk and mental disorders. *International journal of environmental research and public health*, *15*(9), 2028.
7. Ilgen, M. A., Zivin, K., McCammon, R. J., & Valenstein, M. (2008). Pain and suicidal thoughts, plans, and attempts in the United States. *General hospital psychiatry*, *30*(6), 521-527.
8. Rudatsikira, E., Muula, A. S., Siziya, S., & Twa-Twa, J. (2007). Suicidal ideation and associated factors among school-going adolescents in rural Uganda. *BMC Psychiatry*, *7*(1), 1-6.
9. Zivin, K., Eisenberg, D., Gollust, S. E., & Golberstein, E. (2009). Persistence of mental health problems and needs in a college student population. *Journal of affective disorders*, *117*(3), 180-185.
10. Goodwin, R. D., Kroenke, K., Hoven, C. W., & Spitzer, R. L. (2003). Major depression, physical illness, and suicidal ideation in primary care. *Psychosomatic Medicine*, *65*(4), 501-505.
11. Samanta, A., Chowdhury, A., Das, D., Dey, A. K., & Das, M. S. Prediction through machine learning on the dependence of job prospects in the Afro-American community on proficiency in English.
12. Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*, *3*(1), 78.
13. Trupti S. Gaikwad; Snehal A. Jadhav; Ruta R. Vaidya; Snehal H. Kulkarni. "Machine learning amalgamation of Mathematics, Statistics and Electronics". International Research Journal on Advanced Science Hub, 2, 7, 2020, 100-108. doi: 10.47392/irjash.2020.72
14. Aldhyani, T. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H., & Ahmed, Z. A. (2022). Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International journal of environmental research and public health*, *19*(19), 12635.
15. Oppenheimer, C. W., Bertocci, M., Greenberg, T., Chase, H. W., Stiffler, R., Aslam, H. A., ... & Phillips, M. L. (2021). Informing the study of suicidal thoughts and behaviors in distressed young adults: The use of a machine learning approach to identify neuroimaging, psychiatric, behavioral, and demographic correlates. *Psychiatry Research: Neuroimaging*, *317*, 111386.
16. Gee, B. L., Han, J., Benassi, H., & Batterham, P. J. (2020). Suicidal thoughts, suicidal behaviours and self-harm in daily life: A systematic review of ecological momentary assessment studies. *Digital health*, *6*, 2055207620963958.
17. Nusrath Unnisa A; Manjula Yerva; Kurian M Z. "Review on Intrusion Detection System (IDS) for Network Security using Machine Learning Algorithms". International Research Journal on Advanced Science Hub, 4, 03, 2022, 67-74. doi: 10.47392/irjash.2022.014
18. Rajesh P.; Vetrivel Govindarasu. "Analyzing and Predicting Covid-19 Dataset in India using Data Mining with Regression Analysis". *International Research Journal on Advanced Science Hub*, 3, Special Issue 7S, 2021, 91-95. doi: 10.47392/irjash.2021.216
19. Mahalakshmi G.; Shimaali Riyasudeen; Sairam R; Hari Sanjeevi R; Raghupathy B.. "A Survey: Effective Machine Learning Based Classification Algorithm for Medical Dataset". International Research Journal on Advanced Science Hub, 3, Special Issue 9S, 2021, 28-33. doi: 10.47392/irjash.2021.245