Introduction: Good afternoon everyone. First of all ,We are very much thankful to give us the opportunity to share our demonstration . [ I am gourab karak , a under graduate student of Techno Engineering college Banipur , 3rd Year , from CSE department].today on the behalf of my team members(members name) I am going to present our research paper entitled ‘**Prediction of suicidal tendencies using machine learning**’ with the id no 2302026.

Content: in this presentation, we will discuss the details overview of mentioned topic….(all points)

so without wasting much time let’s dive into our main part.

Abstract: now on the behalf of our paper we predict the relationship between suicidal behavior and mental health. As we all know that suicide is a crucial factor in our society mostly depending on mental health. In this research we try to predict suicidal tendency using LR where we also done 10 fold cross-validation, sensitivity, specificity, precision,f1score, and kappa value to verify our model prediction and we come to an accuracy of approx.. 84%.

Paper Introduction: now in the introduction part we have to know what is machine learning. So basically 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, and then test the trained model on different tasks.

According to the World health organization suicide is a third major cause of death among youth at the age level of 15 to 34, and at the same time adults aged 70 and older have the highest rates of suicide deaths in most regions of the world. So our main aim is to identify those people with the suicidal mentality and save their life with the help of our model.

Methodology: : Now I will discuss the methodology part.

First, our job is to collect the data and choose the relevant feature. Therefore we have collected data from the local area of west Bengal about mental health. In this primary data set, We have collected total 97 attributes and 210 data by google form. After we clean the data manually and select our dependent variable **’Thought of self-harm or suicide’** . Then by using the ‘**information gain ratio**’ we will select the most relevant **23** feature as independent variable to make the relationship with the dependent variable**.**

In the next stage, we used a logistic regression model because our dependent variable is binary(0 or 1). So what is logistic regression? Logistic Regression is a numerical method used for binary classification problems, where the target is to analyze one of two possible outcomes. logistic regression is used when the dependent variable has **two possible values**, such as "yes" or "no", "success" or "failure", or "dead" or "alive".

To make a relation among the 23 independent variables with one dependent variable we used linear regression technique and calculate the y\_linear value for each 210 data. After that we put the y\_linear value in sigmoid function and predict y\_logistic value.

After that we trained the model where we split the data set into two parts ‘train and test set’. And with the train set, we train our LR model. We observed that when the threshold value is 0.7 our model work most accurately. So we set the threshold value at 0.7 means if **y\_logistic value is less than 0.7 then our model predicts no which means no suicidal thought is there, else predicts yes.**

after the model is trained we test our model with the test data set and found that our model predicted 84% of data accurately. We discuss the whole result enormously in later slides.

For verifying our prediction, also we have done 10-fold cross-validation and generated a confusion matrix for finding the accuracy, sensitivity, specificity, precision, F1\_score, kappa test.

Flow chart: In this slide, we can see the flow chart of our model. First, we take the data set and run feature selection algorithm to choose the most appropriate features for the dependent variable. Then we split the data set in two part 1st is train set 2nd is test set. After that we trained our model using machine learning with the training data set, then we verify our model with the test data set.

Result: In this slide, we can see the graphical representation. On the left-hand side, we generate the graph for the train test data set where we can see the value of accuracy, sensitivity, specificity, precision,f1score, and kappa as follows for our 50%,66%, and 80% train data respectively. We got an average with the accuracy=**84%**, sensitivity=**77%**, specificity=**85%**, and kappa value =**1**.(100%)

In the right side graph, we have done the same thing for 10-fold cross-validation and we got an average accuracy of almost 83% which is pretty similar to the accuracy of the train test. And the other parameter value is also justifiable.

* At the end of our research we have reached with the accuracy of approx. **84**% for our data set. Not only that our model is able to predict the positive classes accurately approx.. 78%, and true negative classes approx. 86%, and kappa value is 1 which means our model prediction is pretty good. At the last but not least by this research we define some actual reasons for increasing suicide rate and hope our research model should be useful in the field of medical science as a tool and can be able to reduce the reasons behind suicidal thoughts .
* Thank you and have a nice day.

**Q:-What is the difference between MEAN and STANDARD DEVIATION? Which is better among of these?**

ANS:-Standard deviation is more appropriate than Mean.

**Q:- Why we check the CROSS VALIDATION?**

Ans:- To reduce over fitting problem and get the prediction more appropiate, we checked the Cross Validation.

**Q:- What are meaning of all four tuples of confusion matrix?**

Ans:

➢ **TRUE POSITIVE** – means The predicted value was true, and the actual value was also true

* **TRUE NEGATIVE** –It means The predicted value was false, and the actual value was also false.
* **FALSE POSITIVE** – means The predicted value was true, but the actual value was false.

And

* **FALSE NEGATIVE** – It means that actual data and the predicted data both are negative.

**Q:- What are meaning of accuracy , sensitivity, precision, specificity of confusion matrix?**

Ans:

**ACCURACY:** It’s the ratio of the correctly labeled subjects to the whole pool of subjects.Accuracy is the most intuitive one.

**PRECISION:** Precisionmeans out of all the positive predicted, what percentage is truly positive.

**SENSITIVITY:** Recall 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

* **FORMULAS**: -

**ACCURACY = (𝑇𝑃+𝑇𝑁/ ∑𝐾)\* 100**

**RECALL= (TP/FN+TP)\*100**

**PRECISION = (TP/FP+TP)\*100**

**SPECIFICITY = (𝑇𝑁/𝑇𝑁+𝐹𝑃)\* 100**

**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 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.-- **K= P0 - Pe / 1 - Pe**

where, P0 = Probability of agreement, Pe = Probability of random agreement

regression in machine learning:

Machine Learning Regression is **a technique for investigating the relationship between independent variables or features and a dependent variable or outcome**. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes

What is overfitting and underfitting in machine learning?

Underfitting means that your model makes accurate, but initially incorrect predictions. In this case, train error is large and val/test error is large too. Overfitting means that your model makes not accurate predictions. In this case, train error is very small and val/test error is large.

Why is Feature Selection important?

In the machine learning process, feature selection is used **to make the process more accurate**. It also increases the prediction power of the algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones.

What Is Information Gain?

Information Gain, or IG for short, **measures the reduction in entropy or surprise by splitting a dataset according to a given value of a random variable**. A larger information gain suggests a lower entropy group or groups of samples, and hence less surprise.

**Q.What is supervised feature selection?**

**Ans:**

It helps in cutting down the noise in our data and reducing the size of our input data.

Why do we prefer standard deviation over mean deviation?

Standard deviation is considered the most appropriate measure of variability when using a population sample, when the mean is the best measure of center, and when the distribution of data is normal.