Stress Detection by using Machine learning

WEEK 1:- Task
Prerequisite Learning

Prerequisites to learn ML:-

- ➢ Statistics :- it is a discipline concerned mainly with data collection,sorting,analysis,interpretation and presentation.
    - Mean
    - Median
    - Standard deviation
    - Outliers
    - Histogram
- ➢ Probability :- probability describes how likely it is for an event to occur.
    - Notation
    - Probability distribution,joint and conditional
    - Rules of probability is bayes theorem,sum rule and product or chain rule
    - Independence
    - Continuous random variables
- ➢ Linear Algebra :- it is an integral to ML by concept of vector space and matrix operation.
    - Algo in code
    - Linear transforms
    - Notation
    - Matrix Multiplications
    - Tensor and tensor rank
- ➢ Calculus :- it is crucial to building a ML model.
    - Basic knowledge of integration and differentiation.
    - Partial derivations.
    - Gradient or slopes
    - Chain rules for training neural networks.
- ➢ Programming language:- it is good to have a sound foundation in programming as Ml algo.

Some research paper on Stress Detection referring IEEE

- ➢ Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data
- ➢ A Decision Tree Optimised SVM Model for Stress Detection using Biosignal
- ➢ Automatic Stress Detection Using Wearable Sensors and Machine Learning
- ➢ Stress Detection using deep neural networks

Software Development Life Cycle (SDLC) :-

a. Process:- A process is the sequence of steps executed to achieve a goal.
b. Project:- A project is defined by fixed time,scope and resources.

Software Process:- The processes that deal with the technical and management issues of software development are collectively called software processes.

SDLC:- Software Development Life Cycle(SDLC) is a framework that defines the steps involved in developing software at each phase. It covers all the detailed plans for building, deploying and maintaining the software.

Steps of SDLC:-

1. Requirement Analysis
2. Planning
3. Architectural Design
4. Software Development
5. Testing
6. Deployment

Used for SDLC:-
SDLC is to deliver a high-quality product which is as per the customer requirement.

SDLC has defined its phases as requirement gathering, designing, coding, testing, and maintenance. It is important to know that the phase is to provide the product in a systematic manner.

SDLC Model
A software life cycle model is a descriptive representation of the software development cycle.

There are various types of SDLC model used as per project is needed:

- Waterfall Model
- V-Shaped Model
- Prototype Model
- Spiral Model
- Iterative Incremental Model
- Big Bang Model and soon.

In our project Stress Detection we used the waterfall model because it is the first SDLC model used at beginner level. The Waterfall Model is also known as a linear sequential model. In this model the outcome of one phase is the input for the next phase. Development of the next phase starts only when the previous phase is complete. It works step by step. The steps are:-

Software Testing Life Cycle:-

- Requirement Gathering and Analysis:-

  Requirement gathering is an important part to know about our dataset. It collects all the relevant information as per the customer expectation. Data collected from various sources to get valuable data then analyze it.

  Analyzing the data is used to verify the dataset to check any null values or duplicate values or maybe empty values to remove all the null values we used python init after removing the NaN it analyzes the data.

- Design:-

  The requirement gathered in the SRS document is used as an input and software architecture that is used for implementing system development models.

- Implementation or Coding:-

  Once the developer gets the design document the software design is translated into the source code.

- Testing:-

  Testing is an important module to check whether the system is working well or not. Testing is used after completing the coding part. Testers refer to the SRS document to make sure that the software is as per the customer standard.

- Deployment:-

  Once the product is tested, it is deployed in the production environment or first UAT(User Acceptance Testing) is done depending on the customer expectation.

- Maintenance:-

  Maintenance is the last step in SDLC is to deploy a product on the production environment, maintenance of the product from time to time system update or any modification is necessary and taken care by the developer.

Software Testing:-

It is the process of finding errors in the developed product. It also checks whether the real outcomes can match expected results, as well as aids in the identification of defects, missing requirements or gaps.

Importance of software testing:-

- Enhance Product Quality
- Improve Security
- Detect Compatibility with different devices and platforms

Different types of functional testing are:-

- Unit Testing
- Integration Testing
- System Testing
- Smoke Testing
- Interface Testing
- Regression Testing
- Stress Testing
- Load Testing
- Performance Testing and soon

Based on the amount of information you know about the product to test it, software testing can be divided into different types:

- Black Box Testing
- White Box Testing
- Grey Box Testing

Black Box Testing:- In this testing, you have the least amount of information on how the product is built. You don't know the structure of the product, its code and logic.

White Box Testing:- In this testing, you have most information about the product. It is mostly used to make the code better.

Grey Box Testing:- In this testing, you have partial information about the product . this type of testing is helpful to find bugs that the user wouldn't know about.

WEEK:-2 Task
Programming Fundamental(Python)+ Required Installation.

Literature Survey: Prepare below table after reading and analysing IEEE Papers:

Sr.No	Title of the paper	Name of Author	Published Year	Remarks (Methodology and Performance)
1.	Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data	Pramod Bobade Vani M.	2020	WESAD dataset contains data from multiple physiological modalities like three-axis

				acceleration (ACC), respiration (RESP), electrodermal activity (EDA), electrocardiogram (ECG), body temperature (TEMP), electromyogram (EMG) and blood volume pulse (BVP) which is not available in other datasets, which makes this work suitable for the detection of stress in human being. Decision tree has lowest classification accuracy 68.16% and artificial neural network classifier, accuracy has been reached up to 84.32% and up to 95.21% according to WEDSAD dataset
2.	A Decision Tree Optimised SVM Model for Stress Detection using Biosignal	Alana Paul Cruz, Aravind Pradeep, Kavali Riya Sivasankar and Krishnaveni K.S	2020	Eustress and Distress are two kinds of stresses. Eustress is the kind of stress that has positive impact to an individual or it is something which stresses a person but ultimately it make them motivated or energised. distress affects an individual negatively i.e. an individual gets shut down due to stress, anxiety or trauma. The main enhancement of this Tree Optimised SVM model is that it shows improvement in Sensitivity and Elapsed. It is able to generate an accuracy of 96.3%
3.	Automatic Stress Detection Using Wearable Sensors and Machine Learning	Shruti Gedam, Sanchita Paul	2020	Stress can be detected using physical and physiological measures of body. Physical measures include pulse rate, skin temperature,

				humidity, Blood pressure and respiration rate whereas physiological measures can be heart rate, heart rate variability, skin conductance. These can be measured using wearable devices made from low-cost sensors although machine learning algorithms can be used to classify and predict stress level of an individual.There are a diversity of machine learning algorithms which are appropriate for stress detection. Among them Support Vector Machines (SVM), Logistic regression, K-Nearest Neighbour, and Random forest, LR classifiers are give highest accuracy 90% to upto 95.98% are most common.Advances in wearable sensors and mobile computing make it possible to record a variety of physical and physiological signals on a twenty-four hour basis which helps in detection of stress level. Mostly wearable sensor devices like smart band[3], Chest belts[2] are used for data collection.
4.	Stress Detection using deep neural networks	Russell Li , Zhandong Liu	2020-2021	The deep convolutional neural network achieved 99.80% and 99.55% accuracy rates for binary and 3-class classification, respectively. The deep multilayer perceptron neural network achieved 99.65% and 98.38% accuracy rates for binary and 3-class classification,

				respectively.

Data Science is the science of analysing raw data using Statistics and Machine Learning Technique.

Data Science is all about the processes and systems to extract data from structured and semi-structured data. Machine Learning(ML) is a field of AI that allows the software to learn from data to identify patterns and make predictions automatically without any human involvement.

It can further build and train the data model to make real-time predictions.

Machine Learning

In ML we can use various types of algorithms/classifiers used to predict and analyze the raw data in a clean and analysing the data such as random forest, decision tree, logistic regression, SVM and many more.

Before using ML algo we have an idea about python programming used in ML.
ML used python programming language to import used to classify ML algorithms.

Python modules used various types of libraries used to identify and predict the dataset that we are working on.

Some of the python modules libraries are:-

1. Numpy - Numpy was created in 2005 by travis oliphant. It is used for working with arrays. Numpy used “Numerical Python”.

Functions	Used
Type()	Used to build-in python function tell us the type of the object(integer, bool, float)
Index ()	Used for indexing the number
Slicing[ ]	[Start:End]
array_split()	We pass it to the array we want to split and

	the no. of splits

There are few of the functions used in numpy there are many more functions in numpy. Numpy different functions used for analyzing different modules as per requirement.

2. Pandas - Pandas allows us to analyze big data and make conclusions based on statistical theories. It was created by Wes Mckinney in 2008.

Functions	Used
Series()	Column in a table. It holding 1-D array data
Labels ()	Labels with their index
Dictionary[ ]	[Key : Value]
locate()	To locate the rows by loc()
type()	Types of the variables

Pandas is working in Series and DataFrames

Series - it is a 1-D array holding data of any type.

DataFrames - dataframes is working on a multiple rows and columns like 2-D,3-D(whole tables)

Pandas can also be used as Read CSV files

1. A simple way to store a big data set is to use a CSV file.
2. CSV files contains plains text and is a well known format that can be read by everyone including pandas

Pandas Key Note:-

- head() - used to give the first five values of the dataset.
- tail() - used to give the last five values of the dataset.
- info() - used to give information about the dataset and also give information about the number of NaN values.
- to_string() - used to import the whole dataset.

Data Cleaning:- data cleaning means fixing bad data in your dataset

1. Empty cell- to handle empty values there are no. of ways
    - To remove rows

- dropna() used to return a dataframe
- fillna() used to allow us to replace empty cell with some values
- mean() average value
- median() the value in the middle
- mode() the values that appears most frequently

2. Data in wrong format
    - Removes the rows
    - Convert all cell in a column into the same format

3. Data in duplicate
    - duplicate() used duplicate value give as output in a boolean value.
    - drop_duplicate() to remove duplicate value.

Data correlation:- it varies from -1 to 1.
corr() used to calculate the relationship between each column in your dataset . the corr() is the “not numeric” column.

Date & Time

```
Import time
 t=time.time()
 print(t)
```

➔ Current time
```
 localtime = time.localtime(time.time())
```

➔ Calendar
```
cal= calendar.month(year,month)
```

Regular Expression(RegEx)

It is a sequence of characters that defines a search pattern.
re.match() used to find the regular expression

WEEK:-3 Task

ML Specific

MACHINE LEARNING

Machine learning is a branch of AI and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

A computer program which learns from experience is called a machine learning program.

Classification of ML:-

1. Supervised Learning- it is a type of ml used by labeled data.

Both classification and regression problems are supervised learning.

2. Unsupervised Learning- it is a type of learning with unlabeled data. Classification and categorization is not included in this type of learning.
3. Reinforcement learning- it is a type of learning in which it can take action and works as an environment condition. Also known as decision making
4. Semi-supervised learning- it can fall between unsupervised and supervised learning.

ML model of development steps:-

1. Collecting data:- the data can take different resources and the data should be reliable sources.

2. Preparing the data:- the data should be prepared in an evenly distributed manner. It can work according to ETL where E for Extract, T for transform and L for load.

3. Choosing a model:- choosing a model is an important part because the data should be in a structured form and after that it can be split into 2 parts. First 50% of data should be the training part and the remaining should go to the test part .

4. Training the model:- after completing the train and test part the model goes to the training part.it is the most important step in ML model to find patterns and make predictions.

5. Evaluating the model:- for testing purposes for unseen data.

6. Parameter tuning:- after creating and evaluating the process have to improving the accuracy is to increase the size of your dataset(no.of records)

7. Making predictions:- it makes predictions from unseen data of the model.

Machine learning is the process of making systems that learn and improve themselves, by being specifically programmed.

The ultimate goal of ML is to design algorithms that automatically help a system gather data and use that data to learn more. ML is getting systems to think and act like humans, show human-like intelligence and give them a brain.

There are a number of applications that use a ML algo to work automatically without any human interaction. Few application are:-

- Object and image recorigination.
- Detecting fake news
- In a research laboratory
- Self-driven cars
- Automatic Number plate recognition and soon.

NUMPY

Numpy refers to Numerical python. It is an open source library in python that aids in mathematical and numerical calculation and computation.

Numpy is an essential library used to perform mathematical and statistical operations. It is especially suited for multi-dimensional arrays and matrix multiplications.

Importing numpy:-

```
import numpy as np
```

Numpy Arrays

- The array object in numpy is called ndarray, which means an N-dimensional array.

- To create ndarray in Numpy, we use the array() functions.

Numpy Array Functions:-

- ndim():- used to find the dimensions of the array.
- itemsize():- used to calculate the bytes size of each element.
  Each item occupies four bytes.
    - ➢ Dtype:- used to understand the data type of the given element.
    - ➢ Shape:- this array attribute returns a tuple consisting of array dimensions.
    - ➢ reshape():- used to reshape the array
    - ➢ Slicing:- used to extract a range of elements from the array.
- random.rand():- returns a random float between zero and one.
- random.randint():- it takes a size parameter where you can specify the shape of the array.
- mean():- used to compute the arithmetic mean of the given data along the specified axis.
- median():- used to compute the arithmetic median of the given data along the specified axis.
- std():- used to compute the standard deviation along the specified axis.
- append():- used to add new values to an existing array.
- insert():- insert the values in the input array.
- concatenate():- used for joining two or more arrays of the same shape.

PANDAS

Pandas is one of the most widely used python libraries in data science and analytics. It provides a high performance and also easy to use structure and data analysis tools. 2-D tables objects in pandas are referred to as DataFrame as well as Series. It is a structure that contains column names and row labels.

Pandas is well suited for different kinds of data such as ordered and unordered time series data and also used in unlabeled data.

Importing Pandas

```
import pandas as pd
```

Pandas Series
Series is a 1-D array that can contain any type of data.

Basic Operations on Series:-

- Create a series from ndarray
- Create a series from a dictionary
- Accessing data from a series

Pandas DataFrame

A DataFrame is a multi-dimensional data structure in which data is arranged in the form of rows and columns.

Basic Operations on DataFrames:-

- Create a DataFrame from lists.
- Creating a DataFrame from a series dictionary.
- Column selection
- Addition of a new column
- Deleting a column
- Indexing a dataframe with the help of iloc()

Python pandas sorting:-

Pandas can be sort by two ways:-

1. By label :- sort_index() method is used to sort data in pandas.
2. By actual value:- sort_value() method is used to sort the column according to values.

Python pandas GroupBy

GroupBy() is used to

- Splitting the object
- Applying a function
- Combining the result

Bernoulli Naive Bayes

Naive Bayes is a supervised machine learning algorithm to predict the probability of different classes based on numerous attributes.

Naive Bayes is based on the bayes theorem

FORMULA $P(A/B) = \frac{P(B/A).P(A)}{P(B)}$

Where
A: event 1

B: event 2
P(A/B): probability of A being true given B is true - posterior probability
P(B/A): probability of B being true given A is true - the likelihood
P(A): probability of A being true - prior
P(B): probability of B being true - marginalization

Naive Bayes classifiers is based on two essential assumptions:

1. Conditional Independence: all features are independent of each other.
2. Feature Importance : all features are important to get a good prediction and have the most accurate result.

Let there be a random variable ‘x’ and let the probability of success be denoted by ‘p’ and the likelihood of failure be represented by ‘q’

Success: p
Failure: q

q = 1 - (probability of success)
q = 1 - p

i.e.
P(x) = { q = 1 - p x=0
= { p x=1

P(x) = {1 bernoulli trial for success
= { 0 bernoulli trial for failure

WEEK:- 4 TASK
CODING

Bernoulli Bayes algorithm, which is one of the best algorithms for binary classification problems.

Steps for implementing the coding part:-

Step 1:- launch any notebook for importing the code
Step 2:- import the numpy and pandas
Step 3:- import the dataset that we are working on.
Step 4:- check the description of your dataset for better information before working on it.
Step 5:- check if the dataset contains null value or not.
Step 6:- prepare the text column of this dataset to clean the text column with special symbol and language error.
Step 7:- view the most utilized words by individuals sharing about their life issues.
Step 8:- the label column in this dataset contains labels as 0 and 1 where 0 means no stress and 1 means stress.

Step 9:- split the dataset into two part training and testing.
Step 10:- this task is based on the problem of binary classification, we will be using the bernoulli naive bayes algo and also we will see other types of algorithms too to check the accuracy performance.
Step 11:- test the performance of our model on some random sentence based on mental health.

Importing the necessary algorithms and python libraries used in ML to check the result on a sample dataset(spam.csv) file.

In spam.csv file when we split the model the following test accuracy result will be obtain:-

- BernoulliNB Classifier :- 98.70%
- Decision Tree Classifier :- 96.62%
- Random Forest Classifier:- 97.27%

Above result the best fitting model for the spam file is bernoullinb with 98% of accuracy test.

After the sample set we are working on our real dataset ie Stress Management system using ML which is our main aim to get to know the accuracy

We import the python libraries that are used to predict the stress detection model.

WEEK:- 5 TASK
GitHub Deployment

Understanding the git and github process:-

Git is an open-source version control, created in 2005.
Github is a company founded in 2008

When we use git or github it is important to know that:

- We do not need github to use git, but you cannot use github without using git.

Steps to upload a new file in git:-

1. Untracked :- to add the file
2. Unmodified :- edit/ remove the file
3. Modified :- stages the files
4. Staged:- commit the file with a strong message it lied between unmodified and staged to have no other changes into the file.

Steps to create a repository:-

- Create a local git repository

- Add a new file to the repository : after adding a new file in a repository you can check git status update or not by using $ git status.
- Add a file to the staging environment
- Create a commit
- Create a new branch : new branches to make a new feature into a file by using
  Run git checkout-b<my branch name>
- Create a new repository on github : github acts as a collaboratively modify the project code.
- Push a branch to GitHub
- Create a pull request(PR) : it alerts a repository owner that you want to make some changes to their code.
- Merge a PR
- Get changes on github back to your computer.

Also there are some tags used in git:-

1. -A used for add new parameter
2. -m used to commit a strong message
3. Log used to give information about working on a git.

Some importing that we used in our model:-

1. Numpy: used for working with arrays. It also has functions for working in the domain of linear algebra,fourier transform and matrices.

2. Pandas: used for working with dataset.it has functions for analyzing,cleaning,exploring and manipulating data.

3. Matplotlib: used for creating static,animated and interactive visualization.

4. Seaborn: making statistical graphics in python. Seaborn helps you to explore and understand your data.

5. BernoulliNB: it is ML algo used when the dataset is in a binary distribution where the output label is either present or absent.

6. CountVectorizer: it is used to transform a given text into a vector on the basis of the frequency(count) of each word that occurs in the entire text.

7. Natural Language ToolKit: it helps for building python programming that works with human language data for applying statistical NLP.

8. Stopwords: used to eliminate unimportant words, allowing applications to focus on the important words instead.

9. Regression: it helps to predict the outcome of the future event.

10. Wordcloud: word cloud can be made with help of wordcloud library. Word clouds display the most prominent or frequent words in a body of text.

WEEK:- 6 TASK
SUMMARIZATION

Life Cycle of Data Science Project

- Prerequisites for data science
    1. Machine Learning is the backbone of data science
    2. Mathematical Modeling can be extremely helpful to make fast calculations and predictions from what you know of your data.
    3. Statistics is fundamental to data science to extract knowledge and obtain better results from the data.
    4. Computer Programming
    5. Database is the discipline of querying database teaches to understand the query

- Life Cycle of Data Science Project Overview

    1. Define and understand the problem
    2. Data Collection
    3. Data cleaning and preparation
    4. Exploratory Data Analysis
    5. Model building and deployment