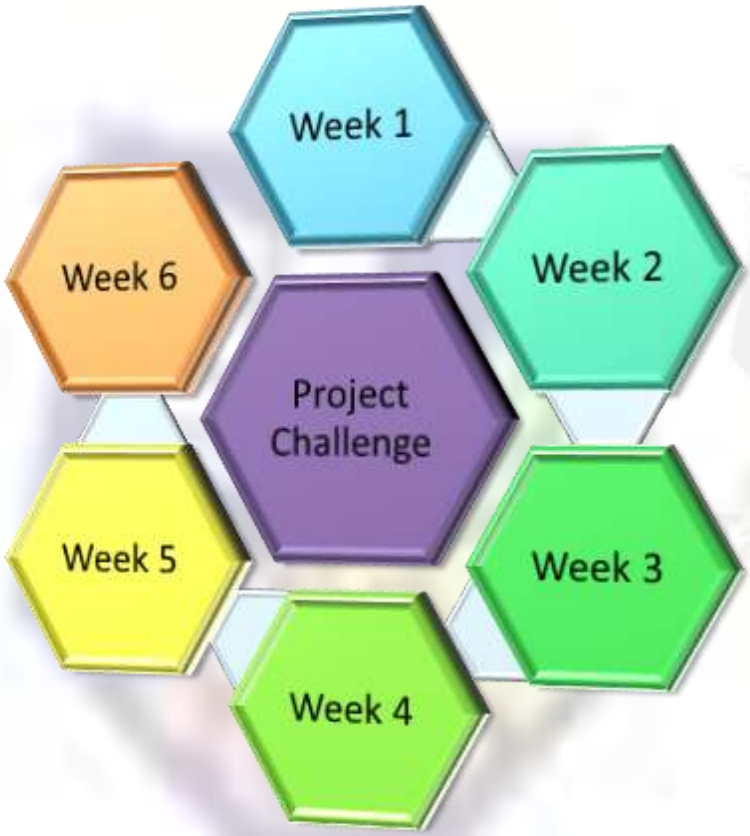
Week 4

Coding



Week 1	Prerequisite Learning
Week 2	Programming Fundamentals (Python) + Required Installation
Week 3	ML Specific
Week 4	Coding
Week 5	Git Hub Deployment
Week 6	Extension + Summarization

Step 1: Launch Jupyter Notebook

Step 2: Import numpy and pandas

Step 3: Import Dataset

Step 4: Check Description of our data

Step 5: Check if dataset contains null value or not.

Step 6: Prepare the text column of this dataset to clean the text column with stopwords, links, special symbols and language errors:

Step 7: View the most utilized words by individuals sharing about their life issues via online entertainment by picturing a word cloud of the text column.

Step 8: The label column in this dataset contains labels as 0 and 1. 0 means no stress, and 1 means stress. We will use Stress and No stress labels instead of 1 and 0.
So let’s prepare this column accordingly and select the text and label columns for the process of training a machine learning model.

Step 9: Split the dataset into training and test sets.

Step 10: This task is based on the problem of binary classification, We will be using the Bernoulli Naive Bayes algorithm, which is one of the best algorithms for binary classification problems.

Step 11: Test the performance of our model on some random sentences based on mental health.

Ex: “I think we need to take care of ourselves”

Ex: “ Sometimes I feel like I need some help”

Bernoulli Naive Bayes

To understand Bernoulli Naive Bayes algorithm, it is essential to understand Naive Bayes.

Naive Bayes is a supervised machine learning algorithm to predict the probability of different classes based on numerous attributes. It indicates the likelihood of occurrence of an event. Naive Bayes is also known as conditional probability.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Naive Bayes is based on the Bayes Theorem.

The Naive Bayes classifier is based on two essential assumptions:-

(i) Conditional Independence - All features are independent of each other. This implies that one feature does not affect the performance of the other. This is the sole reason behind the ‘Naive’ in ‘Naive Bayes.’

(ii) Feature Importance - All features are equally important. It is essential to know all the features to make good predictions and get the most accurate results.

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 Sucesss)
q = 1 - p

Let us consider the example below to understand Bernoulli Naive Bayes:-

Adult	Gender	Fever	Disease
Yes	Female	No	False
Yes	Female	Yes	True
No	Male	Yes	False
No	Male	No	True
Yes	Male	Yes	True

In the provided dataset, we are trying to predict whether a person has a disease or not based on their age, gender, and fever. Here, ‘Disease’ is the target, and the rest are the features.

All values are binary.

We wish to classify an instance ‘X’ where Adult=’Yes’, Gender= ‘Male’, and Fever=’Yes’.

Firstly, we calculate the class probability, probability of disease or not.

P(Disease = True) = $\frac{3}{5}$

P(Disease = False) = $\frac{2}{5}$

Secondly, we calculate the individual probabilities for each feature.

P(Adult= Yes | Disease = True) = $\frac{2}{3}$

P(Gender= Male | Disease = True) = $\frac{2}{3}$

P(Fever= Yes | Disease = True) = $\frac{2}{3}$

P(Adult= Yes | Disease = False) = $\frac{1}{2}$

P(Gender= Male | Disease = False) = $\frac{1}{2}$

P(Fever = Yes | Disease = False) = $\frac{1}{2}$

Now, we need to find out two probabilities:-

(i) P(Disease= True | X) = (P(X | Disease= True) * P(Disease=True))/ P(X)

(ii) P(Disease = False | X) = (P(X | Disease = False) * P(Disease= False))/P(X)

P(Disease = True | X) = (($\frac{2}{3}$ * $\frac{2}{3}$ * $\frac{2}{3}$) * ($\frac{3}{5}$))/P(X) = (8/27 * $\frac{3}{5}$) / P(X) = 0.17/P(X)

P(Disease = False | X) = [($\frac{1}{2}$ * $\frac{1}{2}$ * $\frac{1}{2}$) * ($\frac{2}{5}$)] / P(X) = [$\frac{1}{8}$ * $\frac{2}{5}$] / P(X) = 0.05/ P(X)

Now, we calculate estimator probability:-

P(X) = P(Adult= Yes) * P(Gender = Male) * P(Fever = Yes)

$= 3/5 * 3/5 * 3/5 = 27/125 = 0.21$

So we get finally:-

P(Disease = True | X) = 0.17 / P(X)

= 0.17 / 0.21

= 0.80 - (1)

P(Disease = False | X) = 0.05 / P(X)

= 0.05 / 0.21

= 0.23 - (2)

Now, we notice that (1) > (2), the result of instance 'X' is 'True', i.e., the person has the disease.