APPROACH :

To prepare a classification model using given data set I have used Natural Language Processing (NLP) and Support Vector Machines (SVM) approach.

- NLP is a field in machine learning with the ability of a computer to understand, analyze, manipulate, and potentially generate human language.
- SVM is used because it offers very high accuracy compared to other classifiers.SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points

MODEL INTERPRETATION :

1. Reading and Exploring Dataset

   I have used pandas to easily understand the given data

2. Pre-processing Data

   ➢ **Stemming**

   Stemming helps reduce a word to its stem form. It often makes sense to treat related words in the same way. It removes suffices, like "ing", "ly", "s", etc. by a simple rule-based approach. It reduces the corpus of words but often the actual words get neglected

   ➢ **Removing StopWords**

   Stop words are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, the words like the, he, have etc. Such words are already captured this in corpus named corpus.

   After stemming, Stop words are removed and preprocessed data is stored is copus which was the empty list created

   ➢ **Vectorizing Data**

   Vectorizing is the process of encoding text as integers that is numeric form to create feature vectors so that machine learning algorithms can understand our data

Vectorizing Data: CountVectorizer:

CountVectorizer describes the presence of words within the text data. It gives a result of 1 if present in the sentence and 0 if not present. It, therefore, creates a bag of words with a document-matrix count in each text document. N-grams are simply all combinations of adjacent words or letters of length n that we can find in our source text.

Train Test Split

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm .The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

- **Train Dataset**: Used to fit the machine learning model.
- **Test Dataset**: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

I have splitted the data set, where 80% of data has been given for training and 20% has been given for testing that is test size is 0.2.

Building ML Classifiers: Model selection

- Here I have used Support Vector Machines , which can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.
- The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset.
- The SVM algorithm is implemented in practice using a kernel. A kernel transforms an input data space into the required form. SVM uses a technique called the kernel trick. Here, the kernel takes a low-dimensional input space and transforms it into a higher dimensional space. In other words, you can say that it converts non-separable problem to separable problems by adding

more dimension to it. It is most useful in non-linear separation problem. Kernel trick helps you to build a more accurate classifier.

- Here I have used Liner Kernel, which can be used as normal dot product any two given observations. The product between two vectors is the sum of the multiplication of each pair of input values.
- After building a model its its train and test accuracy score is calculated.
- Models confusion matrix is plotted. A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.
- Its F1 score is calculated

Train and Test Accuracy

- Test Accuracy : 94%
- Train Accuracy : 94%

Limitations Of the Model

- Selecting the right kernel is not an easy task.
- Get slower when dataset size is bigger.
- It classifies through geometry whereas a lot of classification problems probability gives better results.