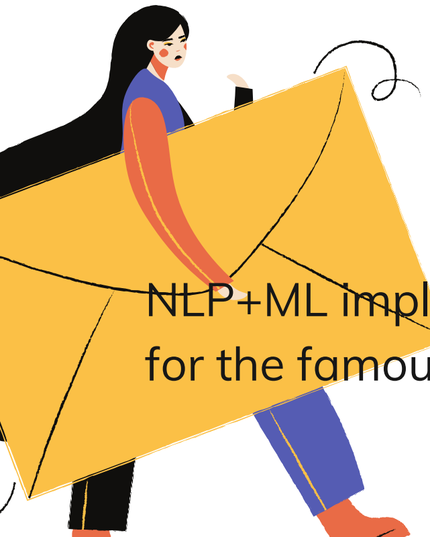
Spam Classifer.



NLP+ML implementation
for the famous topic

The Spam.

Arvind Kumar K
AM.EN.U4CSE19109

Harsha Sathish
AM.EN.U4CSE19123

Rishab Mudliar
AM.EN.U4CSE19136

Project Description.

E-mail has become the preferred medium for communicating official information. Emails provide efficient and effective ways to transmit all kinds of electronic data.

With the rapid increase in email usage, there has also been an increase in the Spam emails. These Spam emails are unsolicited and unwanted junk text, send out in bulk to an indiscriminate recipient list. These emails prevent the user from creating full and sensible use of your time, storage capability and network information measure. It is estimated that spam cost businesses on the order of $100 billion per year

we try to solve it!

The Task

Classify emails as ham or spam using a ML pipeline that uses NLP

Applications.

1. Private companies, who have their own email servers, want their data to be more secure. In such cases, spam classification solutions can be used.

2. Employees of companies need not go through each and every email, if spam filteration is done. Thus, time and work is reduced.

3. Spam filters block viruses from accessing consumer data and prevent any spam mail being forwarded to consumers

4. Spam filters protect the servers from being overloaded with non-essential emails

5. The time and amount of work saved can be used to increase productivity

a Statistical Approach.

1. Statistical NLP comprises all quantitative approaches to automated language processing.

2. Statistical NLP aims to do Statistical Interface for the field of Natural Language

3. Statistical inference in General consists of taking some data and then making some inference about this distribution.

Pipeline.

Here's how we did it

Preprocessing and data visualization

Preprocessing the data using NLP libraries and then visualizing the data we have

Training the data

Training the data on various ML models like SVM, Decision Trees, Random Forest

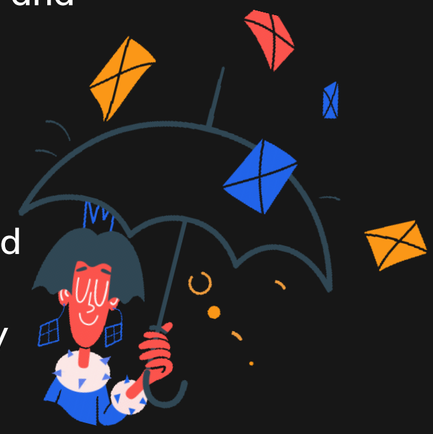
Validation

Comparsion of the algorithms used in training using cross-validation.

data preprocessing

steps involved.

1. Finding and Removing Null Values
2. Removing Punctuations (/,@,%,*)
3. Removing Stopwords
4. Add "Text Length" Column and Lowercasing
5. Tokenization, Stemming and Lemmatization and export processed data to new CSV



CountVectorizer

Convert a collection of text documents to a matrix of token counts.

Tf-IDF Vectorizer

Convert a collection of raw documents to a matrix of TF-IDF features.

Scikit-Learn

Package used for doing the transformations.

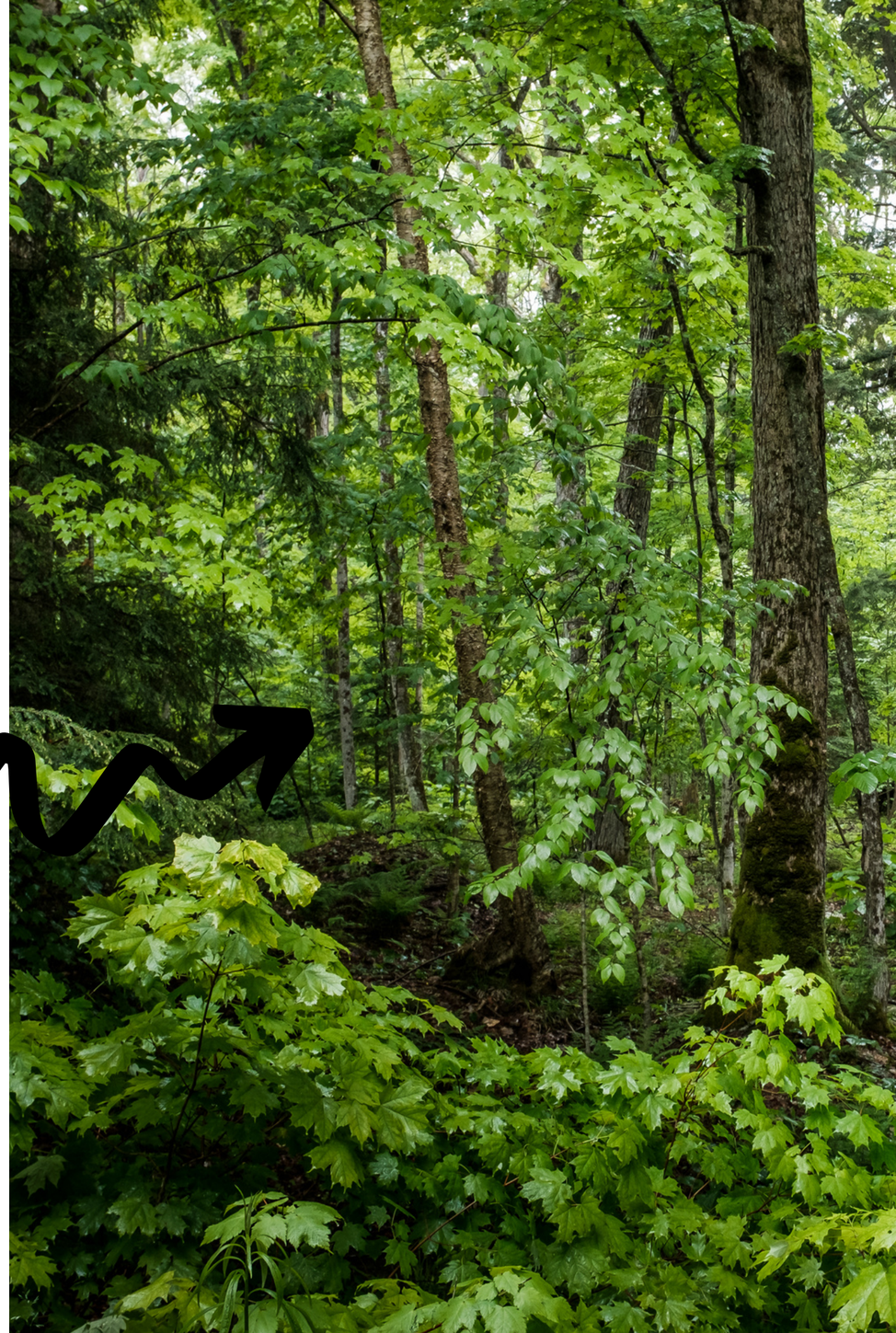
Machine Learning

It is used to process the Natural Language Data and classify the emails as spam or ham

Algorithms used are SVM, Decision Trees, Random Forest

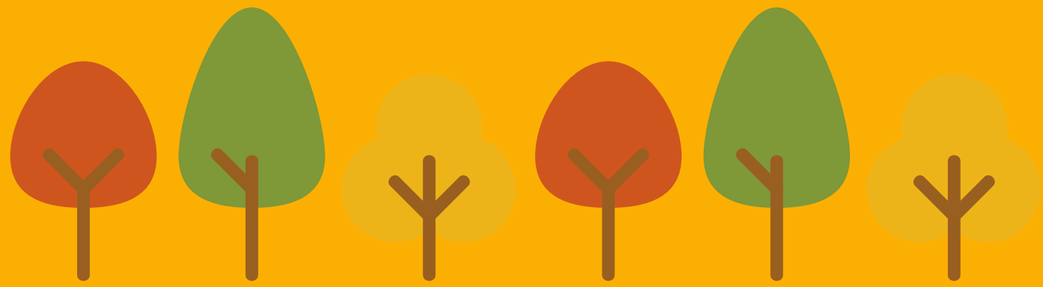
random forest

THE ALGORITHM :P



Decision Tree algorithm belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data)

Learn more here: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html



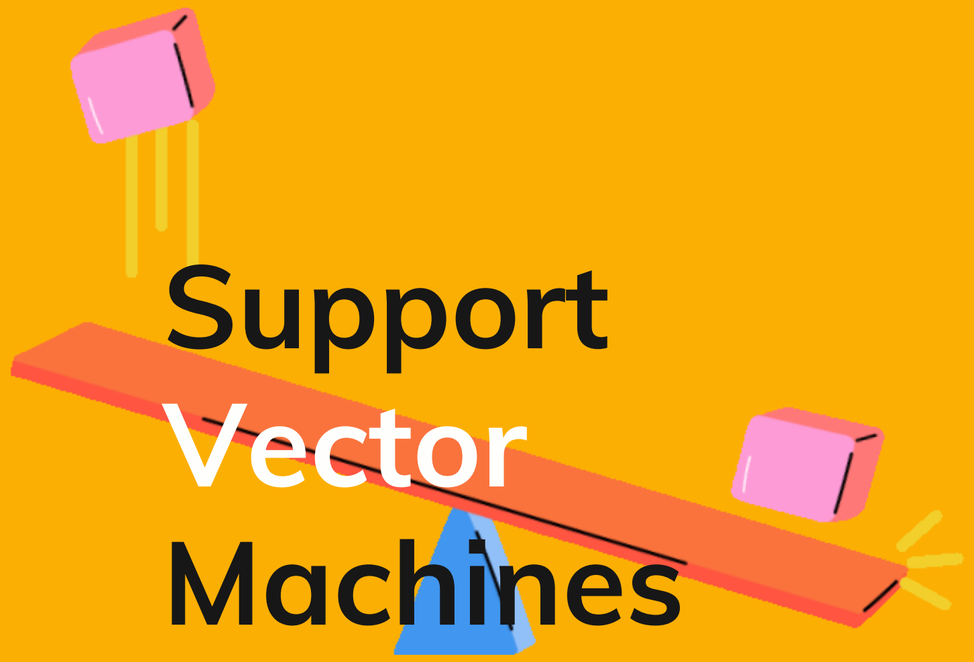
Random Forest is a supervised Machine Learning algorithm, which is used for both classification and regression problems.

Random Forest classifier contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Higher the number of trees in the forest, higher is the accuracy and also prevents the problem of overfitting.

Learn more here: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Support Vector Machines

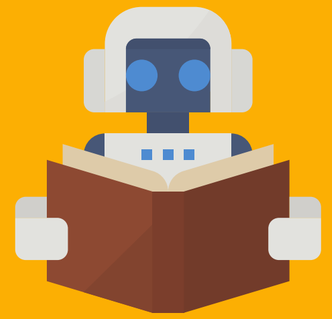


SVM is one of the most popular supervised learning algorithms, which can be used for both classification and regression problems.

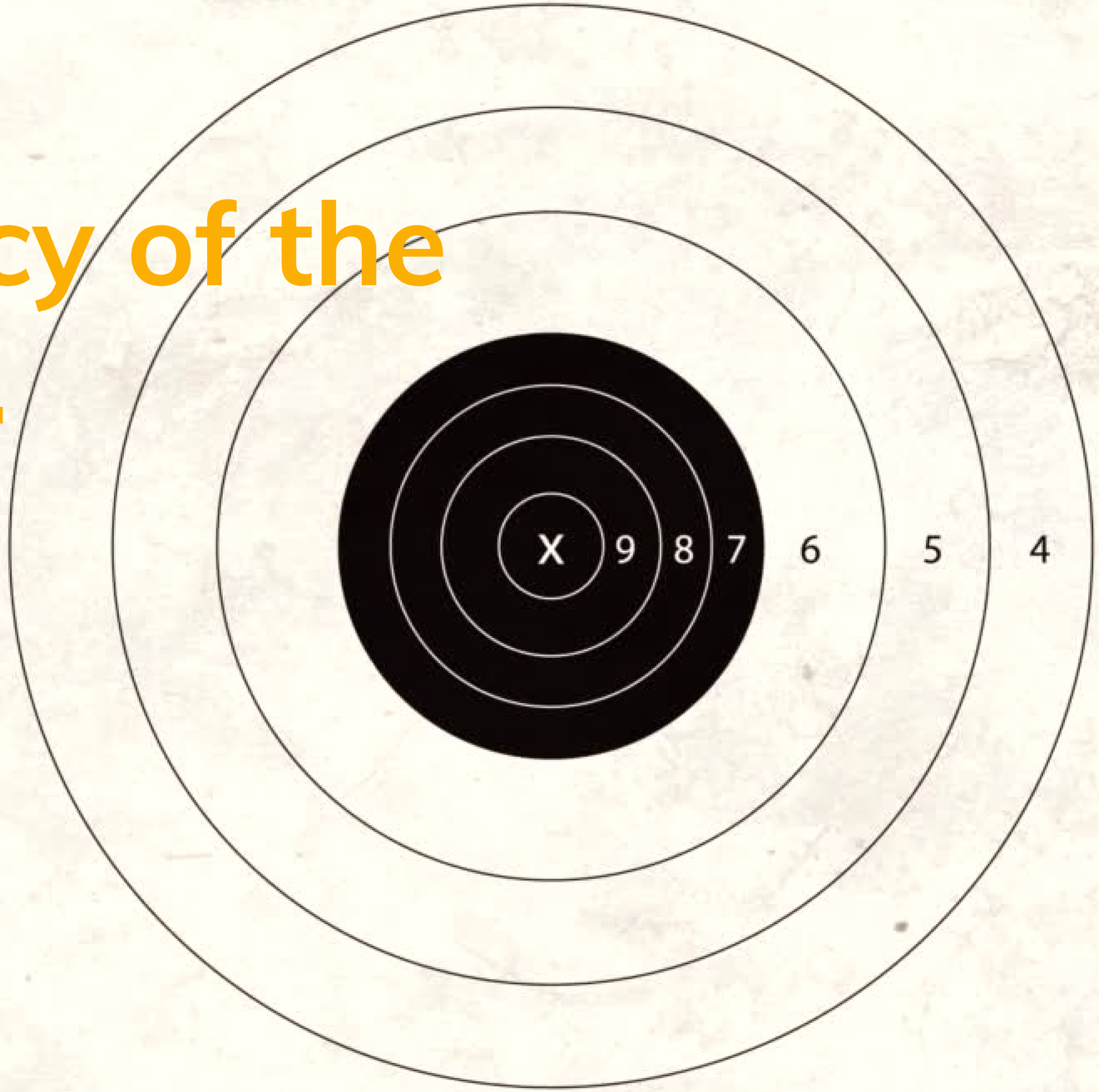
The idea of SVM is to create a line or decision boundary that can segregate the n-dimensional space into classes so that the new data points can be easily put in the correct category.

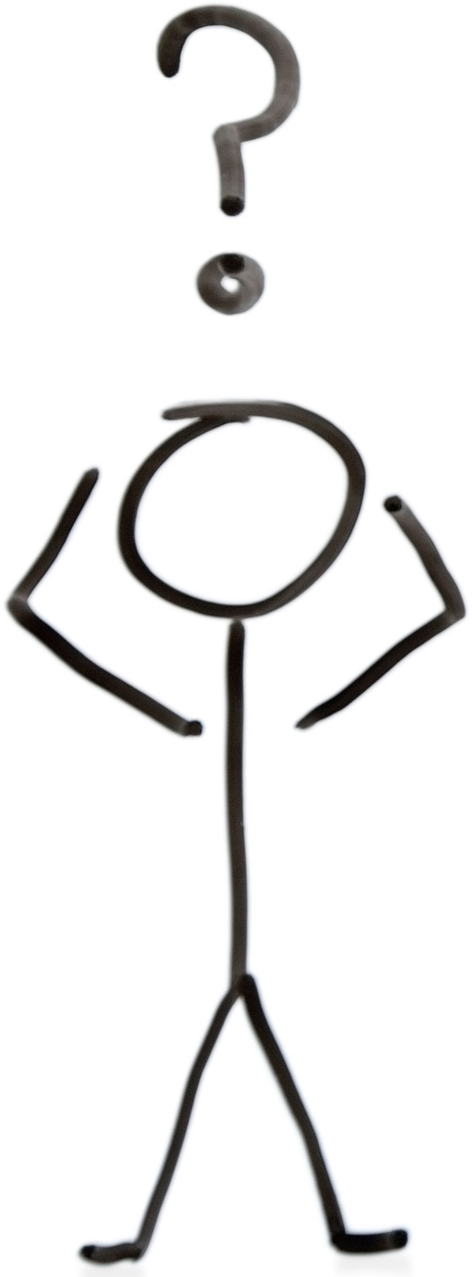
There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but the best decision boundary is to be selected, which is called the hyperplane of SVM.

Learn more here: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

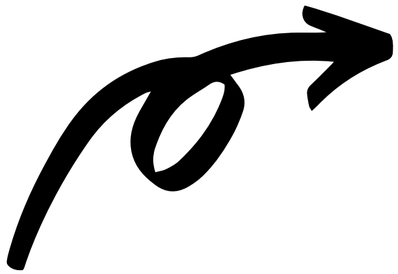


Accuracy of the models.





thank you.



if (any questions == False):