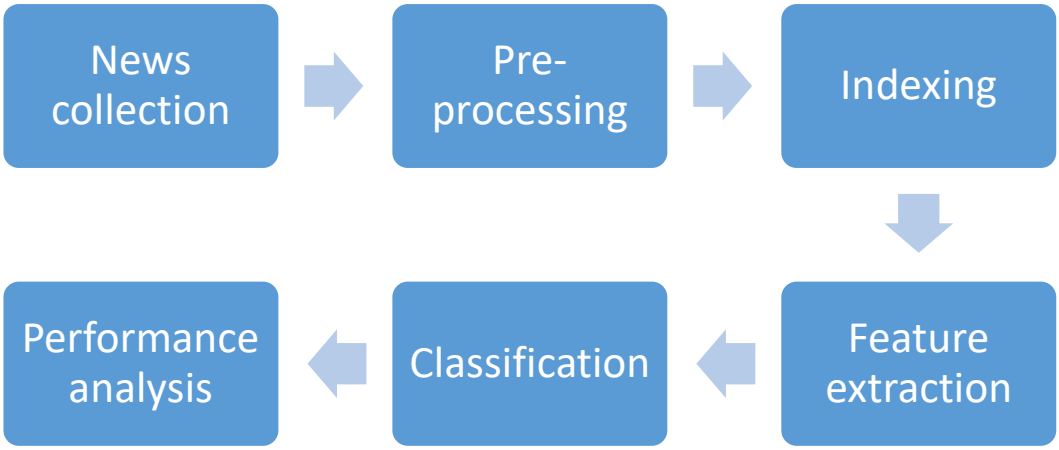
LATEST NEWS CLASSIFIER

Introduction:

Nowadays large amount of information is available in electronic format. So with these data it is necessary to classify and analyse such data and facts which would help in decision making. This is possible because of data mining which is useful in extracting data from huge databases and here it is used to classify the type of news in a given newspaper or article. It is quite challenging as it requires more of pre-processing to convert the unstructured data to some structured information. The categorization/classification of the news benefits the user to access the news of their interest as the number of news increase the difficulty also increases. In this paper the news are been classified based on their content and headlines in a particular article or newspaper.

Workflow:



1. **News Collection:**
   This is the very first news classification is news collection. The news are collected from the various resources like magazines, different types of newspapers, articles etc. and also these files can be gathered in any format like .doc,.pdf etc.

2. **News Pre-Processing:**
   After gathering the news from different sources the next step to be done is pre-processing of that data. To extract the useful information from the data obtained various methods are used like news tokenization, removal of stop words and word stemming.

   - **News Tokenization**: The data obtained is like a whole text and we can't process is directly. So here the pre-processing technique which is news tokenization is used which fragments the whole text into small tokens. Each word in the text is treated as a whole string.

   - **Removal of stop words**: Stop words in the text means the word from which no information is gained like pronouns, conjunction, prepositions etc. also the special characters, semi-colon, full-stop etc. are also the words or

characters from which no information is gained. There are various techniques from which these stop words can be removed. Like the words from where the fewer information is gained are removed. Another one is like list of total 545(approx.) stop words is made and based on that the stop words are dropped.

- **Word Stemming**: This step is used to reduce the word to it's root by removing the suffixes like ed, ing etc. used after the word. There are various types of stemmers like S-stemmer, Porter Stemmer etc.

3. Indexing:

Indexing is the most important step for news classification. Here, the bag of words approach is used to reduce the complexity and difficulty in news classification. Each word in the news content or headline is considered to a vector. Bag of words consist of two things: 1. Vocabulary of known words and 2. Measure of presence of known words and this is used for indexing news headlines. For each word a complete matrix is made as they are in form of vector.

4. Feature Extraction:

When there exist a large number of features and each of the features is a well known descriptive word for each class, a lot of time may be required in classification and it may be possible that expected accuracy may not be achieved and to overcome these issues, a process named as feature selection is adopted in which only those relevant and highly effective features are chosen, which may prove more noticeable for better news classification.

A large number of techniques exists in literature for selecting appropriate features like

- Boolean weighting
- Class Frequency Thresh holding
- Term Frequency Inverse Class Frequency
- Information Gain.

1) Boolean Method Of information retrieval:

The standard Boolean method is the most common and foremost method used for information retrieval. It is based on Boolean logic and set theory concepts in which the data to be searched and query by the user is seen as a sets of terms. Queries are more of a Boolean expression on terms. It uses exact matching scenario for finding the match between the terms to be found in documents. The queries are logically checked by the AND , OR and NOT operators which uses the operators to find the set of documents which contain the "exact words" in the query.

For Instance : x AND y where x and y are elements from the set of terms of query. So the documents which will both x and y will be included in the designated output by Boolean method .

The Boolean model is represented by F, D,Q and R where

F : Boolean algebra over sets of terms and set of documents

D : Set of Indexing terms (keywords) present in the documents that each term is present or not. If the term is present then it is marked as 1 in Set or marked as 0 in Set of indexing terms.

Q : A Boolean Expression which is query needed to be satisfied which consists of related keywords using AND, OR and NOT operators

R : A document is predicted as output which satisfies the query expression.

AND (^) : Intersection of two sets of terms.

OR (v) : Union Of two sets of terms.

NOT (~) : Set inverse

Advantages of Boolean Method :

- Clean and exact formalism
- Easy to implement

Disadvantages of Boolean Method :

- Exact matching leads to too few or too many matching documents.
- As every term is of equal weight , priority classification is not possible.
- As it works on exact matching , more than information retrieval it's more seems like "data retrieval"
- Translation of query in Boolean expression is sometimes difficult.

2) Information Gain :

Information gain measures the probability of particular dataset element occurring to a given value of random variable value. A large value of information gain means that the probability of any data element in this case , Any particular word occurring in the news articles for any particular group. Entropy denotes how much information there is a random variable(keyword) or more precisely it's probability distribution in all the groups.

It is the amount of information gained about a single random variable by observing another variable whose value is known in advanced. But in the case of decision trees which is used here in our case. Information gain means conditional expected value of the univariate probability distribution of one variable to that of another variable.

Information gain is denoted by

$IG(X|Y)=H(X) – H(X|Y)$

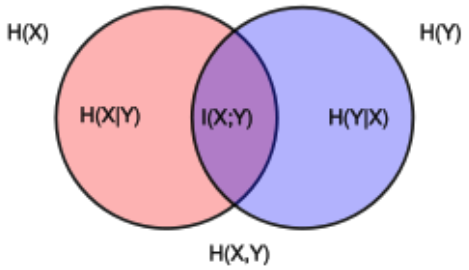
Where,

IG(X|Y) is the mutual information or information gain for X and Y ,

H(X) is the entropy of X

H(X|Y) is the conditional entropy of X given Y

The value of information gain is always greater than zero , higher the value of IG, higher is the relationship between two variables.



Venn representation of Mutual Information

The above figure denotes the additive and subtractive relations among the correlated variables X and Y with their respective entropy.

Uses:

In feature selection and feature transformation in machine learning , used to characterize both the relevance and redundancy of variables

3) **Term frequency inverse class frequency:**

It is denoted by tf-idf , a method to identify how important a word is to classify any particular document in a collection or not , in this case method to classify how any single word determines that particular news article is related to any particular group of news information

The tf-idf value increases proportionally the number of times any particular word is coming in a article and is also considering the articles containing the same so as to find the common words which are repeated more frequently in general. According to a survey held in 2015 , 83% of text-based recommender system uses tf-idf method.

Tf-idf can be used to effectively for stop-words , including text summarization and classification.

Term frequency: it is defined by the number of times any word has occurred in the document , Weight of a term is directly proportional to the term frequency.

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Inverse Document Frequency: If any term is which is a stop word but is occurring excessively in the document which can change the weight and classification of the document so ,An inverse document frequency is used to reduce the weight of the stop words and so emphasize on the relevant keywords which should be occurred is more.

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

$TF\text{-}ICF_w = TF_w * ICF_w$

Where TF is a term frequency of a word “w” and ICF is an inverse

class frequency of a word “w”.

Advantages :

- Easy to compute
- Can remove stop-words effectively
- More efficient than other methods used for the same purpose
- Easy to compute similarity between 2 documents

Disadvantages :

- Used as lexical level , so cannot capture the semantics need

4) Class frequency thresholding:

It gives all classes which are used for classification at least 1 output and afterwards completing the whole classification after seeing some threshold value, if any particular class doesn't have minimum threshold then it is considered redundant or not applicable, which causes a flaw in it with words which have less frequency but more important classification criteria.

5. Classification:

After feature selection the next phase is the classification phase which is an important phase in which the aim is to classify the unseen news to their respective categories. The most common news classification methods are,

- Naive Bayes
- Artificial Neural Networks
- Decision Trees
- Support Vector Machines
- K-Nearest Neighbours

1) Naïve Bayes Algorithm

The naïve bayes algorithm is a statistical classification algorithm suitable for large chunks of data [2].It is the simplest supervised learning algorithm. The first step is to understand the problem and identify the features of the problem and label them. To classify the customers the various features are used for example, their names, age, birth-date etc.

The classification have to phases: 1. Learning Phase and 2. Evaluation Phase [2].

Learning Phase:

In learning phase the model is trained by giving various datasets and train the machine to identify the input given at the particular time.

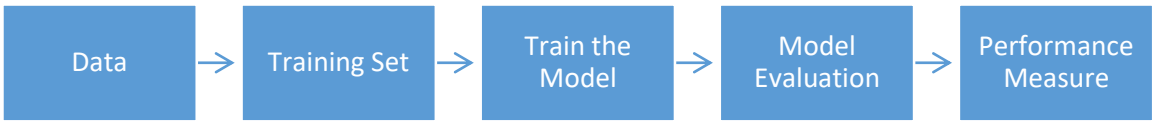


Fig-1 Learning Phase

Evaluation Phase:

In evaluation phase the model is tested by giving input to the machine and check whether it identified correctly or not. This is evaluated on the basis of accuracy, error and precision.

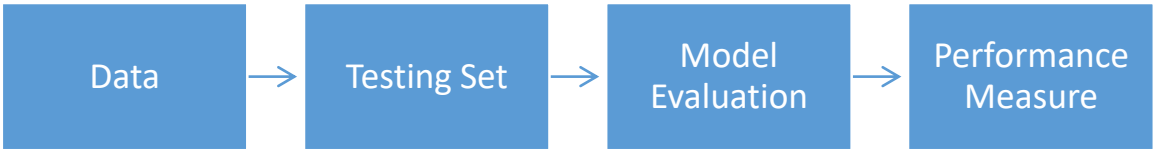


Fig-2 Evaluation Phase

The naïve bayes theorem is based on class conditional independence. Conditional Independence means each and every attribute is independent from each other. This theorem is based on finding the posterior probability, P(class|attribute) from P(class), P(attribute) and P(attribute|class).

Therefore, **Posterior Probability (P(c|a)) = P(a|c) * P(c) / P(a)**

P(c|a) = Posterior Probability of the target class for given attribute

P(a|c) = Probability of the attribute for given class

P(c) = Prior Probability of the target class

P(a) = Prior Probability of the attribute.

To find the Posterior Probability the following steps are followed:

Step-1: Find the probability for each class labels

Step-2: Find the probability of each attribute for each class

Step-3: Put the probabilities in the bayes probability for finding the posterior probability.

Step-4: Observe which class is having the highest probability.

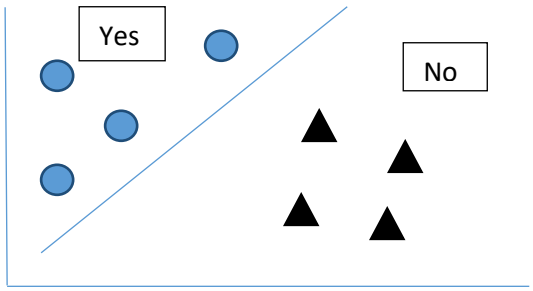


Fig-3 Classification for label yes and no

As It is the classification algorithm, it is used in text classification, sentimental analysis, spam filtering etc.

2) Artificial Neural Networks

Artificial Neural networks works same as neurons works in our brain. The node which is the combination of summation of input and weight associated to it and activation function. The input is given to this node and it generates one output which is equivalent to the action we perform sensed by the neurons of our brain after sensing anything.

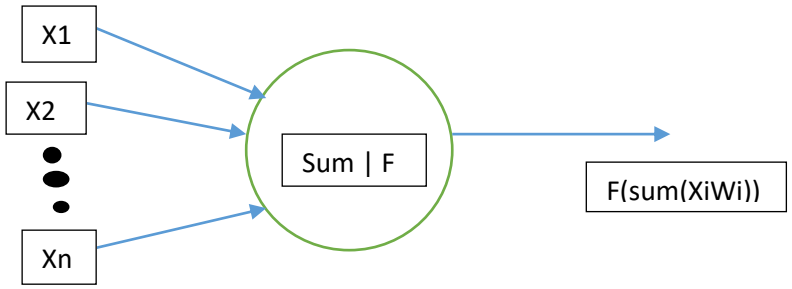


Fig-4 Artificial Neural Networks

Activation Function:

There are 3 types of activation function:

- Linear function
- Heviside function
- Sigmoid function

Linear Function:

As the name suggests this activation function gives the linear relation.

f(x) = a + v

a = bias factor

v = f(sum(XiWi))

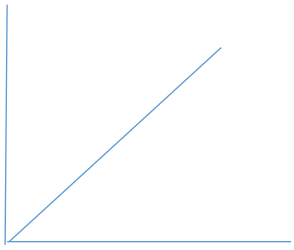


Fig-5 Linear Function

Heviside Step Function:

This function consists of two conditions, 1 or 0.

F(v) = 1 if v>=a

0 otherwise

Here, a is the threshold value.

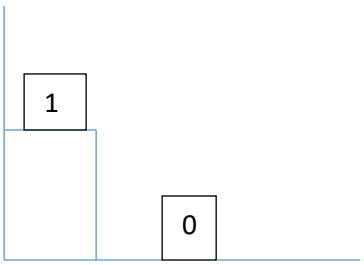


Fig-6 Heviside Step Function

Sigmoid Function:

This function is an exponential function. It means that action is equal to reaction.

F(v) = 1/1 + e^(-v)

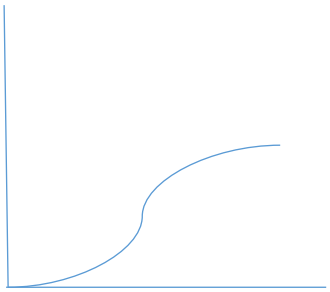


Fig-7 Sigmoid Function

3) Decision Tree

Decision tree is that the supervised machine learning wherever the info is ceaselessly ripping consistent with bound parameters. call tree consists of the common tree elements that are: Nodes, Edges/Branch, Leaf. Nodes square measure the testing worth. the worth of the node is passed to ensuing node or the leaf as a edge or branch. The leaf node predicts the result that represents the category labels or the category distribution.

Types of Decision Tree:

- **Classification trees**
- **Regression Trees**

Classification Trees:

Classification tree square measure of Yes/No varieties. At the end, the choice label square measure either positive or negative. they're build through the method of binary algorithmic partitioning. Binary algorithmic partitioning is that the unvaried method of ripping the info in 2 partitions on every of the branches.

For example, the scholar have done the preparation or not is of Yes/No varieties. Another communicatingple is that if the scholar scored higher than seventy marks in AN exam, this is often additionally a Yes/No varieties call.

Regression Trees:

Regression Trees square measure of continuous information varieties. The target variable or the top of the choice label will be any complex quantity or any continuous values square measure referred to as regression tree.

For example, range of scholars scored higher than seventy marks in AN communicating. Another example is of what percentage students have done the preparation. This takes any continuous worth or a true range.
The most necessary feature of call tree is that the capability of higher cognitive

process information in spite of the length of the provided information. the foundations generated from the choice tree square
measure reciprocally exclusive which implies that no 2 rules conflict
and complete which implies that there's only 1 attribute-value combination for every rule. Order of the foundations doesn't matter.
Decision tree square measure recognized for storing the choice at it's leaf node. So, it's utilized in structured edge detection. the sting structure is keep at the leaf of the choice tree.

4) Simple Vector Machine (SVM)

Simple Vector Machine algorithm is also known as SVM in short, is a supervised learning algorithm that can be used to solve both classification as well as regression types of problems with better accuracy and efficiency [1]. This algorithm is used to find optimal the hyperplane (N – dimensional plane) that distinctly classify the given points perfectly. There are many hyperplanes possible to classify the samples, but support vectors are used to find the hyperplane which has the maximum margin [1]. Support vectors are the data points that are nearer to the hyperplane (minimum perpendicular distance) & which influence the shape and position of hyper plane. If any those data points are removed, then these could change the shape and position of the hyperplane .

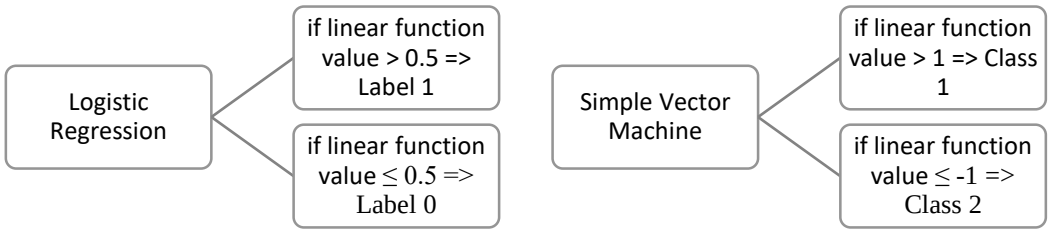


Fig 1. Comparison between linear function in case of Logistic regression & SVM

In this algorithm, our task is to find maximum margin between data points and the hyperplane [1]. The hinge loss function is used to maximize the distance between data points and hyperplane which is defined as follows.

$$H(x, y, f(x)) = \begin{cases} 0 & ; when\ y * f(x) \geq 1 \\ y * f(x) & ; else \end{cases}$$

For the given sample set containing two types of samples: red-cross & blue-triangle as shown in Fig 2. & 3., we are supposed to classify the sample into two groups. There are two feature vectors and a graph are plot for both the samples. Now in Fig 2., we can observe that there are two clusters formed one of red-cross on left side and another of blue-triangles on right side. Here the optimal hyperplane cloud be straight line. Thus, this type of hyperplane is known as linear hyperplane.

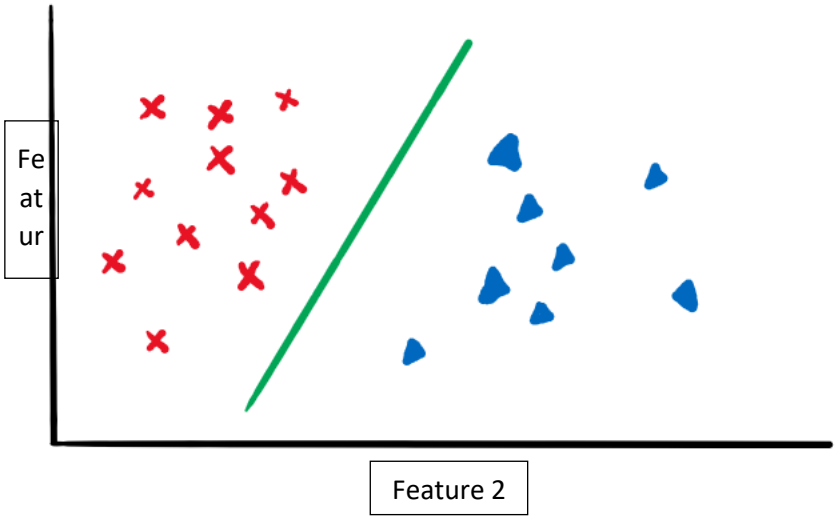


Fig 2. Linear Hyperplane

For the data points which are scattered in the plane in non-linear fashion like one mentioned in Fig 3. could be classified using a non-linear hyperplane like circle, parabola, hyperbola, etc.

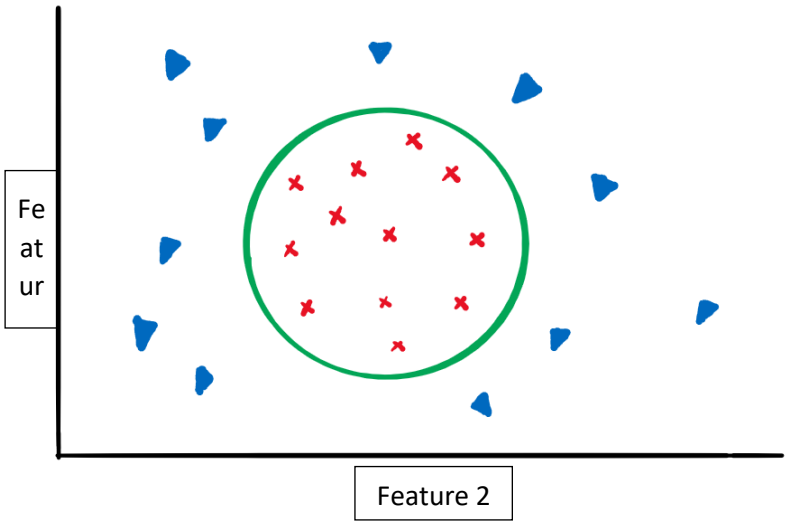


Fig 3. Non-Linear Hyperplane

5) K-Nearest Neighbor (KNN)

K nearest neighbor also known as KNN, is one of the simple supervised learning algorithms that saves all available cases and then classifies the new case based on a measure known as distance function shown in Fig 4.

Euclidean Distance:	$D(X,Y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$

Manhattan Distance:	$D(X,Y) = \sum_{i=1}^{k} \lvert x_i - y_i \rvert$
Minkowski Distance:	$\mathrm{D(X,Y)} = \sqrt[q]{\sum_{i=1}^{k} (\lvert x_i - y_i \rvert)^q}$

Fig 4. Various distance functions used in KNN algorithm.

In the above equations D(X,Y) is the distance between the new case X and the previously stored cases Y, k is the number of neighbors considered and can be any integer.

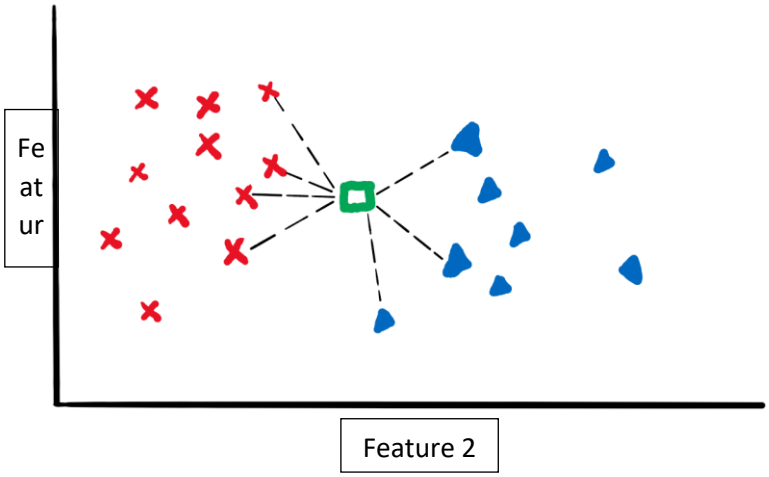


Fig 5. KNN plot

Generally, number of neighbors k is determined by parameters tuning. D(X,Y) is applied with all the previous cases Y and X is added to the cluster which yield minimum distance value of all the cases calculated.

KNN algorithm consists of the following steps:

1. Decide and Initialize number of neighbors be considered k.
2. Calculate the distance.
3. Find the closet distance.
4. Vote for labels.

6. References

[1] Rohith Gandhi 2018, *Support Vector Machine — Introduction to Machine Learning Algorithms*, Towards Data Science, accessed on 8 March 2020, <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

[2] https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn