Original Article

Study Of Supervised Machine Learning Approaches For Word Sense Disambiguation Of Parts-Of-Speech Ambiguity

[1]ARCHANA SACHINDEO MAURYA ,
[2]PROMILA BAHADUR*

[1]Research Scholar, Institute of Technology (IoT), DOCSIS, SRMU, Lucknow.
[2]Associate Professor, Institute of Engineering & Technology (IET), Lucknow.
*Corresponding Author Email ID: promilabahadurpics@gmail.com
DOI: 10.47750/pnr.2022.13.04.302

Abstract

One of the most important applications of Natural Language Processing is Machine Translation (MT). It is an automated process of translation through a computer system. Machine Learning (ML) is one of the recent methods used in MT, and it has become very popular in the area of research over the last numerous years. Ambiguity is a major challenge in MT. ML has given promising results in terms of system learning and predicting results. The text classification technique in Machine Learning is considered as one of the most important methods to resolve Word Sense Disambiguation (WSD). The role of Data set both as Training and Test data is important to predict the required results. We have also done an analysis on supervised machine learning text classification algorithms namely Naïve Bayes', Decision Tree, Support Vector Machine (SVM), K-nearest Neighbor (KNN), Neural Network, Logistic Regression, and Random Forest

Keywords: Artificial Intelligence, Machine Translation, Machine Learning, Hybrid Approach for WSD, Sanskrit Translation, Bayes' Network, Naïve Bayes', SVM, Decision Tree, Random Forest, KNN.

1. Introduction

Machine Learning (ML) is an important aspect of the rising field of Data Science. ML is one of the most exciting and challenging technologies that tends to make computers similar to human beings. We can say that ML makes computers learn. ML is actively being used today in every aspect of life. The term "ML" was defined by Arthur Samuel and he described ML as "a field of study that trains computers to enable an automatic and improving learning process from experience without any human intervention i.e. the capability to learn without being explicitly programmed" [1].

ML is useful in many real-life applications like- web search engines, photo tagging applications, spam detectors, machine translation, etc.

2. Role of Dataset: Training Data and Testing Data

Data sets have a significant role in ML. Data set can be classified as Training data and Test Data. Training Data contains domain-specific information. In the current scenario, Training data contains ambiguous words. For example, as shown in Table 1, the given sentences have "watch" as an ambiguous word. The system is trained to extract features in the vicinity of the given word to predict the POS.

Table 1: Training Data set

Training sentence no.	Training sentences	POS Category
1	This watch is ten minutes fast.	Noun
2	Watch this funny movie today.	Verb
3	I had my watch stolen yesterday	Noun

Test data is a set of sentences that are tested on a given model. The accuracy of prediction is directly proportional to the training provided to the model. The better trained a model is, the better the result is. Test dataset for ambiguous word watch is shown in table 2.

Table 2. Test data set with the prediction of POS

S.No.	Test Sentences	Ambiguous word detected	POS	Predicted POS after applying the Training Model
1.	My watch is five minutes slow.	watch	?	Noun
2.	Watch this movie with your family.	watch	?	Verb

To check the efficiency of different models testing has been done under ten-fold cross-validation. In the ten-fold cross-validation test method, the data set is divided into ten equal parts and in each iteration, nine parts of the given data set are used for the training purpose and the tenth part is used to predict the test data.

3. Ambiguity Resolution and Word Sense Disambiguation (WSD)

In machine translation, data can be translated from the source dialect to the target dialect with the help of machines. Almost all natural languages have reported different kinds of ambiguities. When we translate from one language to another, these ambiguous words need to be disambiguated properly for the appropriate translation into either source or target language. The ambiguity problem can be solved with a disambiguation process. Machine Translation (MT) is the most important application in which we use the approaches of WSD for the removal of different kinds of ambiguities [2]. Therefore, WSD is the disambiguation method for selecting the correct meaning of an ambiguous word.

WSD is frequently defined as the process of determining the POS and specific meaning/right sense of an ambiguous word in a given sentence. When a system deals with a word that has multiple meanings, this technique identifies the contextual meaning of that sentence. Therefore, ambiguous words have several meanings and the WSD technique can help to select the exact meaning of that word [3]. There are various approaches that can be applied to the WSD process. These approaches are knowledge-based, supervised learning, and unsupervised learning approaches. The approaches are shown in Fig. 1.

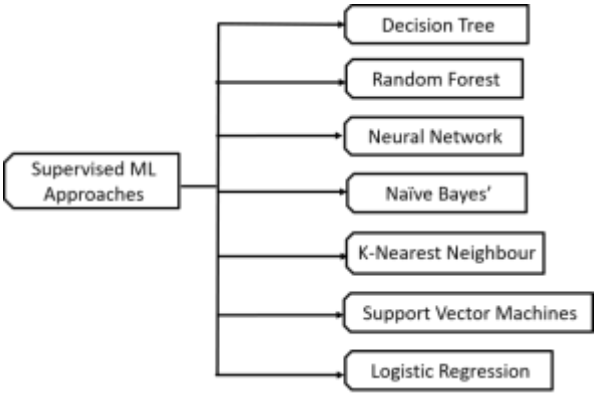

Fig. 1. Classification of Supervised Machine Learning Approaches

3.1 Decision Tree Learning

Decision tree learning is a tree-based method and it is one of the most important and widely used supervised learning algorithms. These algorithms provide great accuracy and stability to predictive models. A tree is a data structure that is frequently used to store data for searching and sorting processes. Furthermore, these trees can be used to make decisions and implement systems based on these decisions. The basic architecture of the binary decision tree is characterized by the following points:

1. Each binary tree is denoted by the internal nodes and the leaf nodes.
2. Always the conditions can be implemented by the internal nodes and they evaluate to “true” or “false”.
3. All internal nodes have two sub-trees:
   a. Left sub-tree
   b. Right sub-tree
4. The first node of the tree is called the root node and this node does not have any parent.
5. There is a parent node for all leaf nodes and internal nodes.
6. The leaf node does not have any children.
7. These binary trees are used to make quick decisions.

Let us consider the sentences of training data in Table 3. The decision tree is shown in Fig. 2.

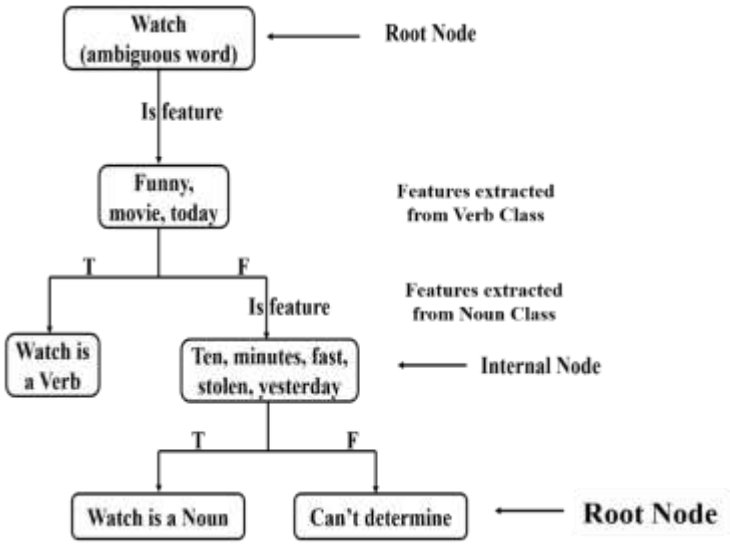

Fig. 2. The Decision Tree of sentences in Table 3

3.2 Random Forest Algorithm

Random Forest (RF) is also one of the powerful and widely used methods for data exploration, data modeling, and predictive modeling. This classifier is an ensemble algorithm in which various decision trees are grouped together. An ensemble decision tree will have a low variance and a high accuracy value in comparison to a single decision tree.

A random forest can be built by using the decision trees for the same data set, but the trees cannot be correlated. The result of this algorithm will be in the form of a tree and constructed from the results of separate decision trees [4].

Each decision tree contributes votes for the-segmentation of a new item based on its qualities or attributes, and the tree with the most votes is chosen for segmentation. The visual illustration of the random forest approach for the training data set watch is shown in Fig. 3.

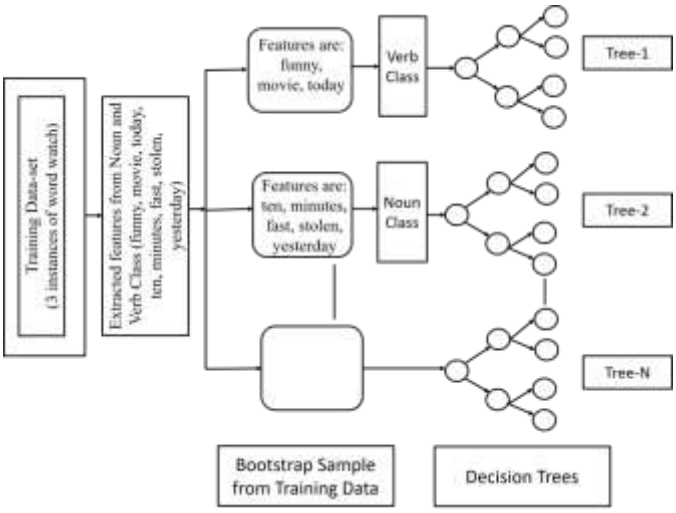

Fig. 3. Visual representation of Random Forest Algorithm for the table 3 sentences

3.3 Neural Network Learning

The brain is one of the most complicated organs in the human body and the main task of the brain is to learn the new things. Neural networks are computing devices, that are parallel to each other and they are capable of attempting the brain functions. The main objective of this learning method is to develop a model of the brain so that a system can perform the computational tasks much faster than the current system. These computational tasks include segmentation, data clustering, pattern recognition, optimization etc.

3.3.1 Artificial Neural Network Learning

Artificial Neural Networks (ANNs) are one of the most important and significant features of neural networks because they have the ability to learn like the human brain [5]. ANN is made up of the processing units that are called neurons. A neuron has input unit called dendrites and output units called synapses or axons.

An artificial neural network consist of three layers i.e. input layer, hidden layer and, an output layer. The first layer is called the input layer and it contains neurons that transfer information to the middle hidden layer. The hidden layer dose the computations on the received information and transfer the calculated information to the output layer. The Neural Network created for Table 3 is shown in Fig. 4.

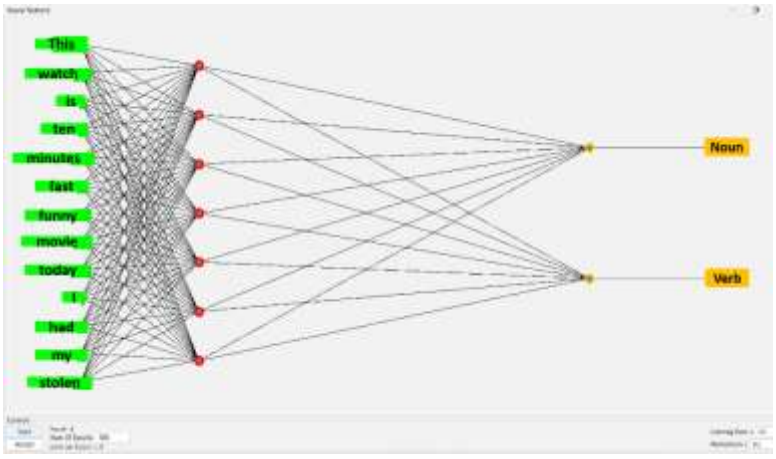

Fig. 4. Screenshot of Neural Network created for Training data set Sentences

3.4 Bayesian Learning

Bayesian learning method is one of the popular method for the document categorization. This is the probability based method and provide the quantitative approach for evaluating the performance of other algorithms. In Bayesian learning, the conditional probability of an occurring event is calculated and that event is also correlated to some other events [6].

3.4.1 Bayesian Network

A Bayesian network is a directed acyclic graph in which nodes represents variables and edge reflect conditional dependencies. It is a probability-based graphical model that uses a directed acyclic graph to express a set of features and their conditional relationship [7]. This network has two parts:

a. Directed acyclic graph
b. Conditional probability table

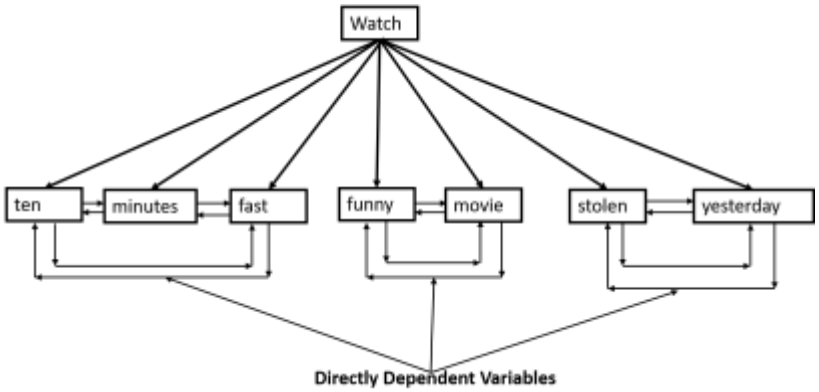

Fig. 5. Bayesian network of Training Dataset

(i) Bag-of-word model: This is the most common method for text classification. BoW model is a simple way of extracting features from the text documents. BoW model records the total number of a particular word with in a document. In the other terms the BoW can also be represented by the histogram on the basis of independent variables [8].

Example:

Sentence1: "This watch is ten minutes fast".
Sentence2: "Watch this funny movie".
Sentence3: "I had my watch stolen yesterday".

Based on the text documents, a group of words is produced for each sentence by tokenizing the sentences to produce a dictionary of the words, as follows:

Sentence1: "This", "watch", "is", "ten", "minutes", "fast".
Sentence2: "watch", "this", "funny", "movie".
Sentence3: "I", "had", "my", "watch", "stolen", "yesterday".

Bag of word for each sentence is represented as follows:

BoW1: "This=1", "watch=1", "is=1", "ten=1", "minutes=1", fast=1".
BoW2: "watch=1", "this=1", "funny=1", "movie=1".
Bow3: "I=1", "had=1", "my=1", "watch=1", "stolen=1", "yesterday=1".

The resultant BoW of the above three sentences will be the union of BoW of these three sentences i.e.

BoW = Bow1 U BoW2 U BoW3

Therefore the final BoW will be represented as follows:

BoW="This=2", "watch=3", "is=1", "ten=1", "minutes=1", fast=1", "funny=1", "movie=1", "I=1", "had=1", "my=1", "stolen=1", "yesterday=1".

(ii) Collocation: A sequence of two or more consecutive words in a document or in a sentence is termed as collocation. These words has syntactic and semantic features [9]. In the collocation model, the feature vector of the set of a selected word is at the specific position located to the left or right of the target word that is ambiguous in nature. The collocation feature vector can be extracted from the window size two, it means two words to the left and two words to the right from the target word [10]. For example, consider the following training sentence:

"This watch is ten minutes fast".

In this sentence, our target word is "watch" that is at position n and assuming window size of +/- 3 from the target word after removing stop words, the selected words will be:

"This watch ten minutes fast".

This classification technique is based on the Bayes' theorem and is mostly used for text classification.

3.4.2 Bayes' Theorem

Another name for Bayes' theorem is Bayes' Rule or Bayes' Law, which is used to determine the conditional probability of an event with the help of prior probability. Bayes' Theorem is widely applied in the field of machine learning.

Suppose there are two events A and B, then the Bayes' theorem formula can be given by the following equation:

$$P(\frac{A}{B}) = \frac{P(B/A)*P(A)}{P(B)} \ldots\ldots\ldots\ldots\ldots (5)$$

Here,

A and B are the events.

P(A | B) is the Posterior Probability of event A after event B has occurred.

P(A) is the Prior Probability before the event occurs.

P(B | A) is the Likelihood Probability of event B after occurring evidence of event A.

P(B) is the Marginal Probability

To find out conditional probability feature in a given class Bayes' rule is applied [11- 12]. This algorithm helps to calculate the conditional probability of each value of a term and feature in a given sentence. The highest value will result in the most appropriate result.

3.4.3 Naïve Bayes' Algorithm

Naïve Bayes' algorithm is a supervised probability-based machine learning algorithm that works on Bayes' theorem. This algorithm is mostly used for document segmentation problems. There are many variables present in the training data set and these variables are independent of each other. These independent variables are called "features." To compute the likelihood of certain features in a given class, this approach uses Bayes' theorem [13-15]. In this context, the uppermost value shows the most "significant" class. A Naïve Bayes model is easy to implement and useful for high-dimensional training data sets i.e. with a high number of rows and columns. It is simple and easy to implement and also outperforms all other classification methods.

The following formula can be used to compute the probability value for each class:

$$C = \underset{c}{\operatorname{argmax}} \quad P(c) \prod_{x=1}^{m} P(\frac{fx}{c}) \ldots\ldots\ldots (8)$$

$$C = \underset{c}{\operatorname{argmax}}\, P(c/f1, f2, \ldots \ldots \ldots fx) \ldots\ldots (6)$$

$$C = \underset{c}{\operatorname{argmax}} \frac{P\left(f1,f2,\ldots\ldots\frac{fx}{c}\right)P(c)}{P(f1,f2,\ldots\ldots fx)} \ldots\ldots\ldots\ldots (7)$$

Here, C represents the POS of the ambiguous word w

f1, f2, …….fx represents the selected feature

x the total number of retrieved features.

The overall prediction process of Naïve Bayes' Classification approach is shown in Fig. 21.

3.5 Instance-Based Learning: K-Nearest Neighbor (KNN)

With the instance-based learning approach, the system can generalize new instances based on some shared properties after learning from the training data. Another name for instance-based learning is memory-based learning, or lazy learning. The time complexity of this learning algorithm depends on the size of the training data set. The most popular instance-based learning algorithm is K-nearest Neighbor (KNN).

K-Nearest Neighbors (KNN) method is a non-parametric grouping method that is basic but active in many situations [16]. This method saves all available cases and categorizes new ones based on the votes of its k neighbors [17]. KNN is a popular statistical method for segmentation and is used for unlabeled observations after assigning them to the class. Features of observations are collected for the training data set and the testing data set. The algorithm can be applied to the segmentation and regression issues. Two important concepts can be implemented in this algorithm:

a. One strategy is based on calculating the distance between two similar characteristics in the new and training samples. Find the nearest k neighbors first, then decide the category to which the neighbor belongs, and finally determine the category of the new sample [18].
b. Another approach is to choose the value of k, which determines how many neighbors the KNN algorithm can use. The number of k that is chosen correctly has a substantial impact on the KNN algorithm's performance [19].

It is very essential to determine the optimum value for k in this method. It should not be too small and also not too large. If the value of k is too small, the model will be overly particular and will not generalize well, making it susceptible to noise. The model provides good accuracy on the training data set but poor prediction on the testing data set in this case. This condition is called the overfitting of the model. If the value of k is selected as too large, then the generalized model will not be a good predictor of both the training and testing data sets. This condition is called under-fitting [20].

The Euclidean distance between two points can be calculated using the formula below:

$$= \sqrt{((X_2 - X_1)^2 + (Y_2 - Y_1)^2)} \ldots\ldots\ldots (4)$$

Here,

X1, and Y1, are the coordinates of the test word.

X2, and Y2 are the coordinates of the matching feature.

3.6 Support Vector Learning

A support vector machine is like a sharp knife and it works on both small and complex data sets. It is a stronger and more powerful algorithm for building machine-learning models. It is mostly used to solve categorization challenges. We can draw each data element as a [21] point in n-dimensional space with a value for each characteristic using this technique. After that, we can perform segmentation by finding the hyperplane. This hyper-plane clearly differentiates the two classes [22].

The primary objective of this strategy is to differentiate between negative and positive situations by using a larger margin. The margin is the distance between the hyperplane and the closest positive or negative points. The two positive and negative points, which are the best parallels to the hyperplane, are called supportive vectors.

3.7 Logistic Regression

This is a probability–based method supervised text segmentation approach. It is a predictive algorithm. This algorithm is used when the data sets are unconditional and the output will be in binary form. The segmentation problems based on the binary output are known as binary segment problems [23].

Information gathered from the training data set is shown in Table 3.

Table 3: Gathered Information for Training data set and testing dataset

Kappa statistics	-0.5
Mean absolute error	0.6667
Root mean squared error	0.8165
Relative absolute error	133.3331
Root relative squared error	157.5674

The logistic regression curve generated from table 3 data is shown in Fig. 6.

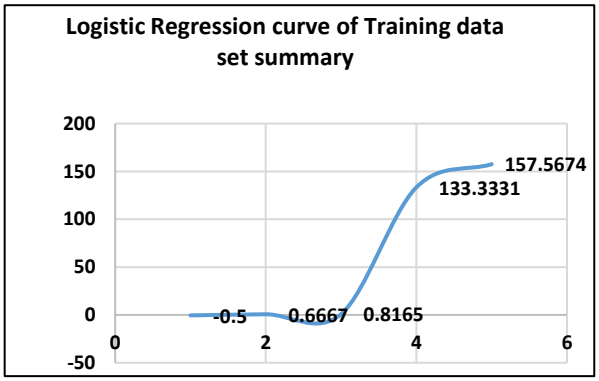

Fig. 6. Logistic Regression Curve

4. Conclusion

Ambiguity has been an open challenge in the field of Machine Translation. Ambiguity refers to words having multiple meanings, senses, POS, etc. The central idea of the paper is to study the machine learning approaches to solve the parts-of-speech ambiguity problem during machine translation process. provide a solution to ambiguity issues using existing Machine Learning Algorithms. Different ML algorithms have been tested in terms of accuracy and efficiency on the pre-processed data set. The test was carried out on the machine learning tool WEKA. The study will be extend in future for the larger dataset.

References

1. Samuel, Arthur L. (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 44: 206 226. CiteSeerX 10.1.1.368.2254.doi:10.1147/rd.441.0206
2. Bahadur, P., & Chauhan, D. S. (2014, August). Machine Translation—A journey. In 2014 Science and Information Conference (pp. 187-195). IEEE.
3. Navigli, R. (2009). Word sense disambiguation: A survey. ACM computing surveys (CSUR), 41(2), 1-69.
4. Venkatesan, N., & Priya, G. (2015). A study of random forest algorithm with implementation using weka. International journal of

innovative research in computer science and engineering, 1(6), 156-162.
5. Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5), 717-727.
6. Book Machine Learning Tom M. Mitchell
7. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2), 131-163.
8. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications, and challenges. In 2019 International Engineering Conference (IEC) (pp. 200-204). IEEE.
9. Wermter, J. (2008). Collocation and term extraction using linguistically enhanced statistical methods (Doctoral dissertation, Jena, Univ., Diss., 2008).
10. Kumari, S., & Singh, P. (2013). Optimized word sense disambiguation in Hindi using genetic algorithm. International Journal of Research in Computrer & Communication Technology, 2(7), 445-449.
11. https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/
12. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.
13. Szepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning, 4(1), 1-103.
14. http://www.cs.upc.edu/~escudero/wsd/00-ecai.pdf
15. Nyein Thwet Thwet Aung, Khin Mar Soe, Ni Lar Thein, "A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011, pp. 1-7.
16. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.
17. Wang, L. (2019, December). Research and implementation of machine learning classifier based on knn. In IOP Conference Series: Materials Science and Engineering (Vol. 677, No. 5, p. 052038). IOP Publishing.
18. Cuong Anh Le and Akira Shimazu, "High WSD accuracy using Naive Bayesian classifier with rich features", PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo, pp. 105-114.
19. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).
20. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.
21. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
22. S.N. Jeyanthi, "Efficient Classification Algorithms using SVMs for Large Data sets," A Project Report Submitted in partial fulfillment of the requirements for the Degree of Master of Technology in Computational Science, Supercomputer Education and Research Center, IISC, BANGALORE, INDIA, Jun. 2007
23. https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_quick_guide.htm#