A Comparative Study of Classification Algorithms: Social Network Ads

Abstract-Social media advertising is a market that did not exist just one and a half decade ago, but the rapid increase in popularity of social networking platforms is projected to generate a $37 billion in revenue by 2020.Social advertisements offer the opportunity to build awareness, target based on geographic, demographic and behavioural data. Also, many of the ads can be contextually placed adjacent to relevant topics. Many of the platforms also offer remarketing opportunities for visitors who abandoned the site or shopping cart and returned to the social network. This paper analyses the categorical dataset from kaggle to determine whether a user purchased a particular product from a Social Network ad.

**Introduction –** Classification is a supervised machine learning approach to a datasets, where we categorize the data into a desired and distinct number of classes where we can assign labels to each class. There are numerous applications for classification in multiple domains like medical diagnosis, email spam prediction, credit card approval, etc.

There are two types of learners in classification as lazy learners and eager learners.

Lazy learners-They simply store the training data and wait for a testing data to appear. When it does, the classification is conducted based on the most related data within the stored training data. Compared to eager learners, lazy learners have less training time, but take more time in predicting.Ex. k-nearest neighbour.

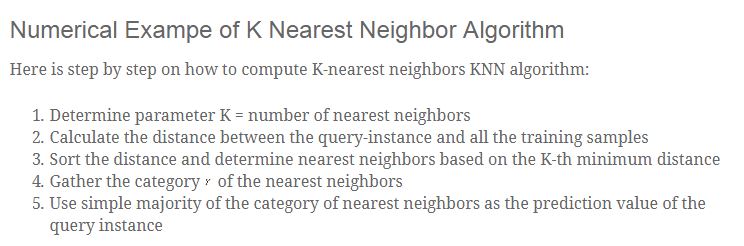
Eager learners-They construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and lesser time to predict.Ex. Decision Tree.

There are a lot of classifications and algorithms available now, but it is not possible to conclude which one is superior to other. It depends on the application and nature of the data set on which it is applied. Hence, we are performing this comparative study of selected algorithms on the Social Network Ad dataset.

**ALGORITHMS**

**1.KNN ALGORITHM**

<https://people.revoledu.com/kardi/tutorial/KNN/KNN_Numerical-example.html>



**Pros and Cons of KNN**  
 **Pros**

1. Easy to understand
2. No assumptions about data
3. Can be applied to both classification and regression
4. Works easily on multi-class problems

**Cons**

1. Memory Intensive / Computationally expensive
2. Sensitive to scale of data
3. Not work well on rare event (skewed) target variable
4. Struggle when high number of independent variables

**2.SVM ALGORITHM**

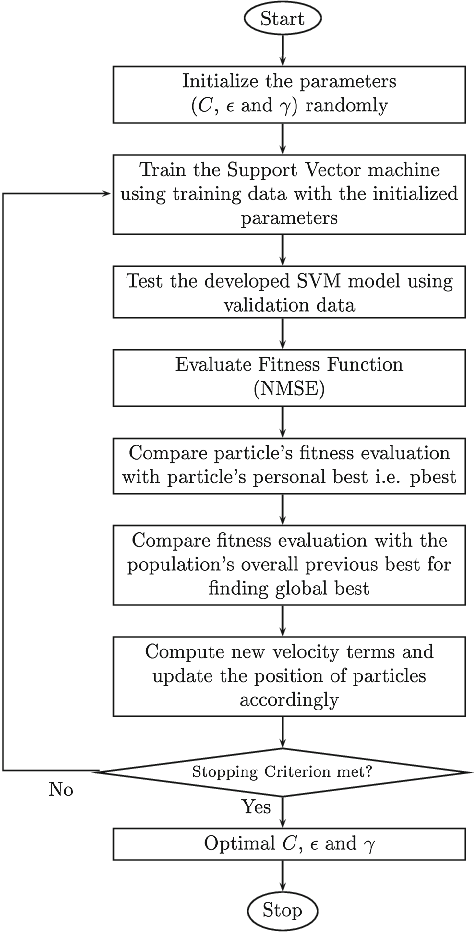
**Pros & Cons of Support Vector Machines**

**Pros**

* Accuracy
* Works well on smaller cleaner datasets
* It can be more efficient because it uses a subset of training points

**Cons**

* Isn’t suited to larger datasets as the training time with SVMs can be high
* Less effective on noisier datasets with overlapping classes



**Algorithm:**

**Steps:**

1.start

2.initialize the parameter randomly c,€,Ƴ randomly

3. training the svm using traing data with the initialized parameters

4.test the developed svm model using validation data

5. evaluate fitness function

6.compare particles fitness evaluation with particles personal best

7.compare new velocity terms and update the position o f the particles accordingly

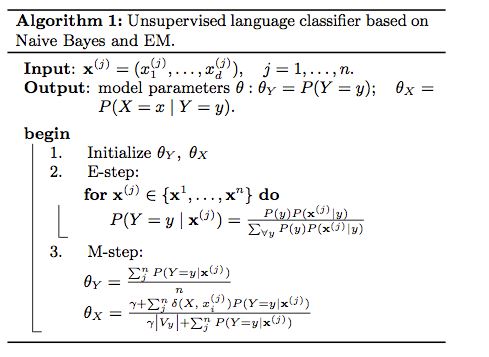
8.stopping criterion met?

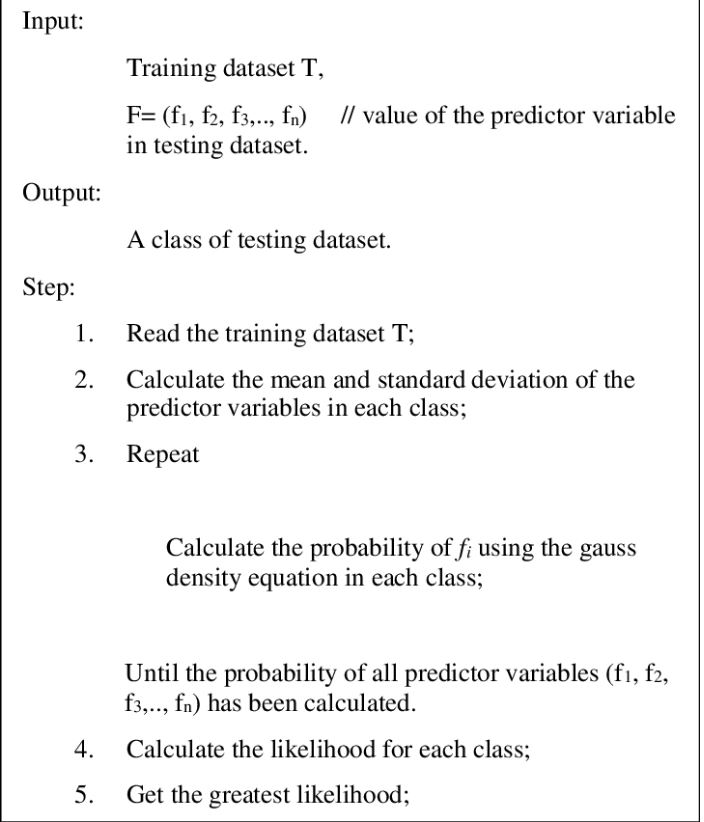
If no goto step 3

9. if yes optimal C,€,Ƴ

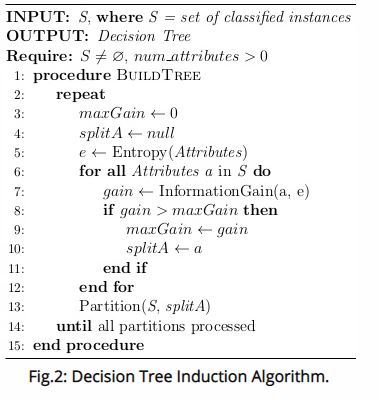
10.stop

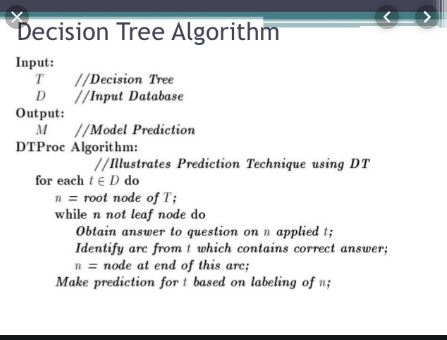
**3.NAIVE BAYES**

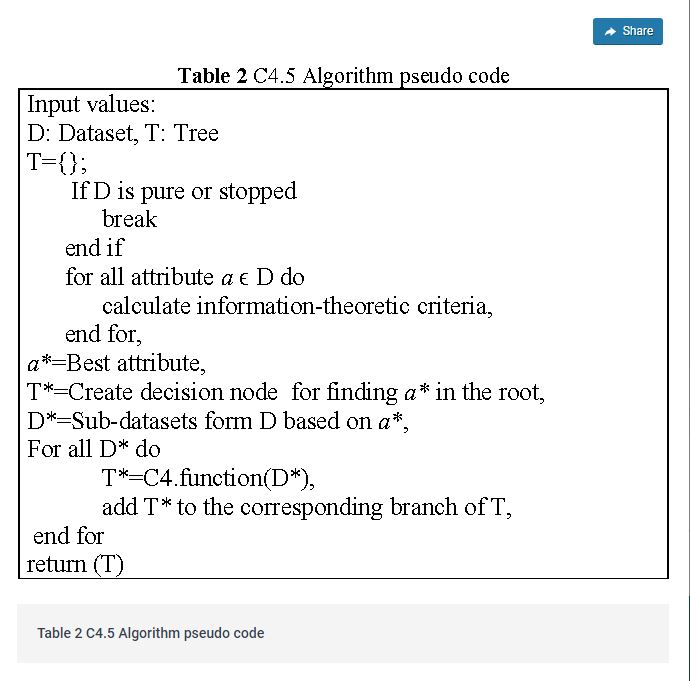
****

****

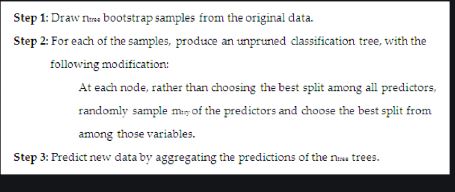
**4.DECSION TREE ALGORITHM**

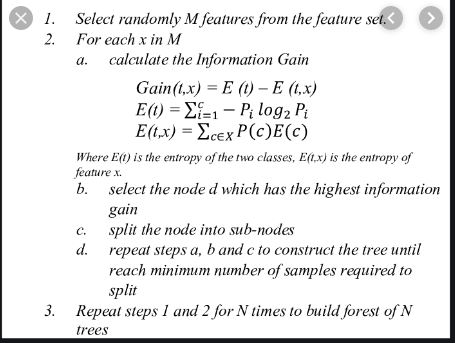
****

****

****

**5.RANDOM FOREST ALGORITHM**

****

****