Data Mining: An experimental approach with WEKA on UCI Dataset

Ajay Kumar
Dept. of computer science
Shivaji College
University of Delhi, India

Indranath Chatterjee
Dept. of computer science
Faculty of Mathematical sci.
University of Delhi, India

ABSTRACT
Data mining became a popular research field these days. The reasons that attracted attention in information technology, the discovery of meaningful information from large collections of data. Data mining is the perception that we are data rich but very much information poor. Large amount of data is available all around but we can hardly able to turn them in to useful information. The comparative analysis of available classification and clustering algorithms is provided in this paper through theoretical and practical approach with WEKA tool. It also includes the future directions for researchers in the field of data mining.

Keywords
Data Mining, Weka, Classification, Clustering, UCI dataset.

1. INTRODUCTION
Data mining is the study of patterns which are hidden in data that is not easily visible. We use interestingness criteria in data then find out hidden pattern successfully

Type of data:

- Tabular (Relational, multidimensional): Transaction data, etc
- Spatial: Remote sensing data, etc
- Temporal: Log information
- Streaming: Multimedia data, Network traffic
- Spatio-temporal: GIS
- Tree: XML data
- Graphs: website, activity log
- Text, Multimedia.

Types of Interestingness.

- Frequency
- Consistency
- Rarity
- Correlation
- Length of occurrence
- Periodicity
- Abnormal behavior
- Graphs: website, activity log

2. CLASSIFICATION TECHNIQUES.
Classification is a machine learning technique used for prediction of group membership for data instances. Here in this paper, we have presented some of the basic classification techniques. Several important kinds of classification methods including decision tree induction, k-nearest neighbour classifier, Bayesian networks, case-based reasoning, fuzzy logic techniques and genetic algorithm.

DECISION TREE INDUCTION
Decision trees are techniques that classify instances by sorting them based on dimension values. Each of the nodes in a decision tree represents a particular feature in an instance to be classified and each of the branches represents a particular value which can be assumed by the node. Data are classified starting at the root node and then sorted accordingly on their feature values.

The basic algorithm for the decision tree is a greedy algorithm that constructs the trees in a top-down recursive way of divide and conquer manner.

The algorithm may be summarized as follows.

1. Create a node N
2. If the samples are all of the same class C, then
3. Return N as a leaf node which is labelled with the class C
4. If the attribute list is empty, then
5. Return N as a leaf node which is labelled having the most common class within the samples
6. Select the test attribute, the attribute among the attribute list having the highest information gain
7. Label the node N with the test attribute
8. For each of the known value a_i of the test attribute
9. Grow a branch from node N where the condition test attribute = ai
10. Let si be the set of samples where test attribute = ai
11. If si is empty then
12. Attach a leaf labelled having the most common class in samples
13. Else attach the node returned by function which generate decision tree

BAYESIAN NETWORKS:
Bayesian Network is a graphical model to find the probability relationships among certain set of features. This network structure 'S' is a directed acyclic graph and the nodes in 'S' are in the one-to-one mapping with the features 'X'. The arcs represent the casual influences within the feature sets while the lack of possible arcs in 'S' encodes without any conditional dependency.

K-NEAREST NEIGHBOR CLASSIFIERS:
The nearest-neighbour classifier is based on the method of learning by analogy. N dimensional numeric attributes are used to describe the training samples. Each of the samples represents a particular point in an n-dimensional data space. Same way, all the training samples are placed in an n-dimensional data space. When an unknown sample is given, a

k-nearest neighbour classifier searches for the data space for 'k' training samples that are nearest to the unknown sample. "Nearness" or "closeness" can mainly be defined in terms of 'Euclidean distance' between the two data points, where Euclidean distance of the two data points X & Y can be stated as, $X=(x_1, x_2, \ldots, x_n)$ and $Y=(y_1, y_2, \ldots, y_n)$ is

$$d= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The K-nearest neighbor classification algorithm popularly known as KNN is one of relatively simple method in data mining classification techniques, study and research KNN classification algorithm for data classification and data mining technology has a very important significance. An attribute Weighted KNN(W-KNN) method is utilized for reducing the irrelevant attributes. And the weight parameter can distinguish the different effective of different attributes in classification. K Nearest Neighbor (KNN) algorithm is a classification and prediction method used widely in pattern recognition and data mining, and it is a supervised machine learning method.

3. CLUSTERING TECHNIQUES

Clustering is a technique in which a given data set is divided into groups called clusters in such a manner that the data points that are similar lie together in one cluster. Clustering plays an important role in the field of data mining due to the large amount of data sets.

There are various clustering algorithms available like DBSCAN, CLARA, CURE, CLARANS, k-Means etc.

Clustering algorithms are classified according to:

- The type of input
- Clustering criterion defining the similarity between the objects
- Concepts on which clustering analysis techniques are based likewise fuzzy theory, categorical data and numerical data.

PARTITIONAL CLUSTERING

Partitional clustering tries to decompose the data set directly into a set of disjoint clusters. The criterion function that the algorithm tries to minimize by assigning clusters to the peaks in the probability density function or the global structure may emphasize the local structure of the data. The global criteria mainly involve the minimizing some of the measure of dissimilarity in the samples within each cluster, and thus maximizing dissimilarity of the different clusters too.

Time Complexity: O (nkl) where n is the number of patterns, k is the total number of clusters and l is the total number of iterations. Generally, the value of k and l is fixed prior to attain the linear time complexity.

Space Complexity: O (k+n) and addition storage is required for storing the data.

HIEARCHICAL CLUSTERING

A dendogram is constructed which is the tree of clusters, depending on the medium of proximity in Hierarchical technique. Each cluster node contains other nodes called child nodes and the nodes from same parent are called sibling nodes. Hierarchical techniques have a property of quick termination. Example of Hierarchical clustering include: CURE, BIRCH, CHAMELEON etc.

NEURAL NETWORK

Neural networks are non-linear data modeling tools that simulate the working of brain. They are used to identify the relationship between the patterns depending on the input and output values.

FUZZY CLUSTERING

Fuzzy clustering is a reasoning based technique in which associations among the patterns and the clusters is done on the basis of membership functions. Fuzzy clustering generates overlapping clusters.

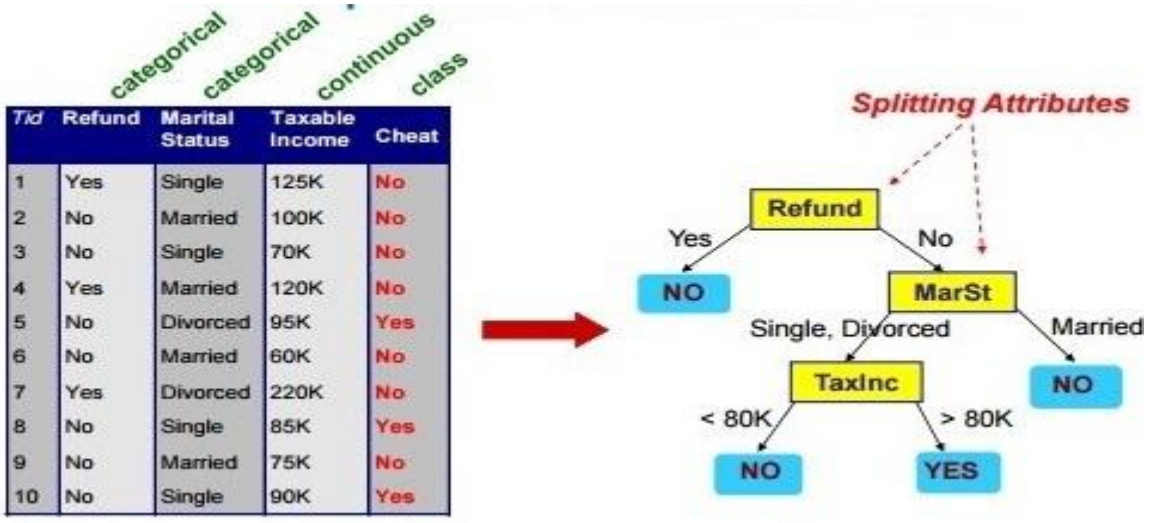

Fig. 1. Decision Tree model

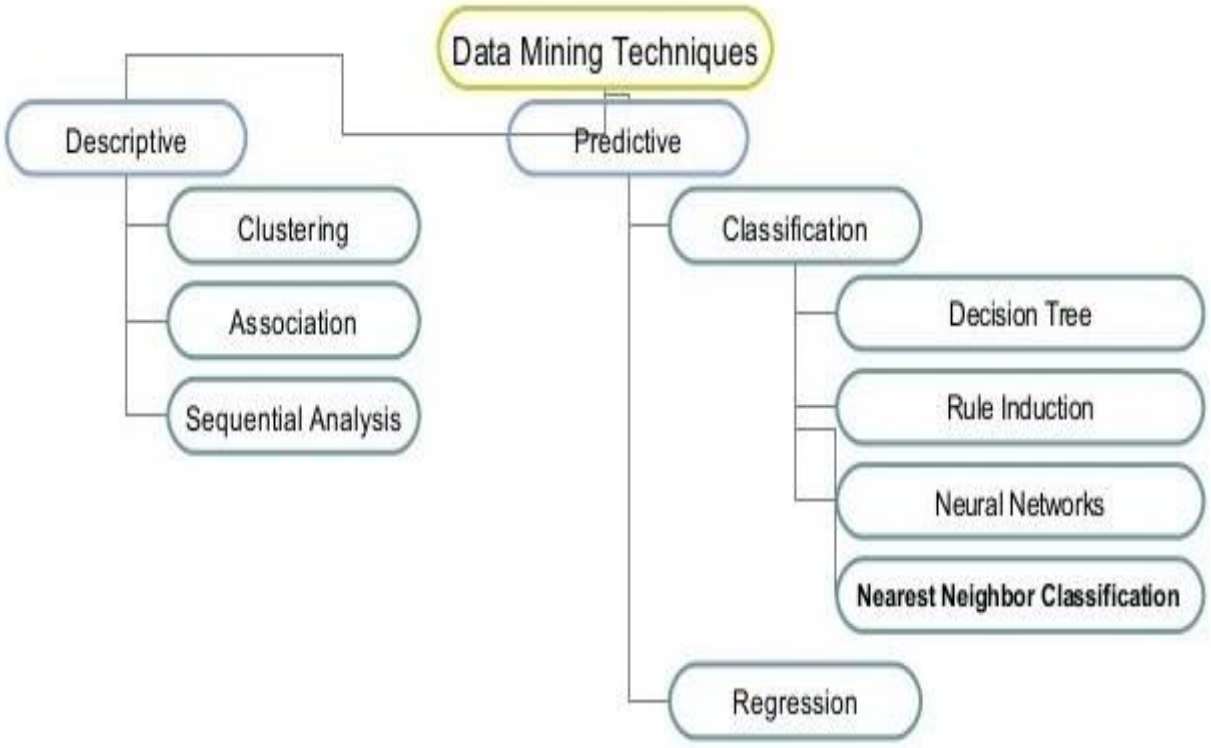

Fig. 2. Data Mining Techniques

4. WEKA: DATA MINING TOOL

Weka is a data mining tool which includes a large variety of funcions. Its GUI is easy to use and makes it accessible to all users, while its flexibility and extensibility is also user friendly.

It is written in Java and released under the GNU General Public Licence (GPL). It can be run on Windows, Linux, Mac and other platforms

EXPERIMENT

Iris dataset.

The Iris data set, a small, well-understood and known data set. The typical task for the Iris data set is to classify the type of iris based on the measurements. It is one of the most analyzed data sets in statistics, data mining, and multivariate visualization.

iris.arff

Relation: iris

No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
1	5.1	3.5	1.4	0.2	Iris-se...
2	4.9	3.0	1.4	0.2	Iris-se...
3	4.7	3.2	1.3	0.2	Iris-se...
4	4.6	3.1	1.5	0.2	Iris-se...
5	5.0	3.6	1.4	0.2	Iris-se...
6	5.4	3.9	1.7	0.4	Iris-se...
7	4.6	3.4	1.4	0.3	Iris-se...
8	5.0	3.4	1.5	0.2	Iris-se...
9	4.4	2.9	1.4	0.2	Iris-se...
10	4.9	3.1	1.5	0.1	Iris-se...
11	5.4	3.7	1.5	0.2	Iris-se...
12	4.8	3.4	1.6	0.2	Iris-se...
13	4.8	3.0	1.4	0.1	Iris-se...
14	4.3	3.0	1.1	0.1	Iris-se...
15	5.8	4.0	1.2	0.2	Iris-se...
16	5.7	4.4	1.5	0.4	Iris-se...
17	5.4	3.9	1.3	0.4	Iris-se...
18	5.1	3.5	1.4	0.3	Iris-se...
19	5.7	3.8	1.7	0.3	Iris-se...
20	5.1	3.8	1.5	0.3	Iris-se...
21	5.4	3.4	1.7	0.2	Iris-se...
22	5.1	3.7	1.5	0.4	Iris-se...
23	4.6	3.6	1.0	0.2	Iris-se...
24	5.1	3.3	1.7	0.5	Iris-se...
25	4.8	3.4	1.9	0.2	Iris-se...
26	5.0	3.0	1.6	0.2	Iris-se...
27	5.0	3.4	1.6	0.4	Iris-se...
28	5.2	3.5	1.5	0.2	Iris-se...
29	5.2	3.4	1.4	0.2	Iris-se...
30	4.7	3.2	1.6	0.2	Iris-se...
31	4.8	3.1	1.6	0.2	Iris-se...
32	5.4	3.4	1.5	0.4	Iris-se...
33	5.2	4.1	1.5	0.1	Iris-se...
34	5.5	4.2	1.4	0.2	Iris-se...
35	4.9	3.1	1.5	0.1	Iris-se...
36	5.0	3.2	1.2	0.2	Iris-se...
37	5.5	3.5	1.3	0.2	Iris-se...
38	4.9	3.1	1.5	0.1	Iris-se...
39	4.4	3.0	1.3	0.2	Iris-se...
40	5.1	3.4	1.5	0.2	Iris-se...

Fig. 3. IRIS dataset from UCI datacenter J48 classification

We have applied a decision tree model called J48 on the IRIS dataset would allow us to predict the target variable of a new dataset record. Decision tree J48 is the implementation of algorithm by the WEKA project team.

CLASSIFICATION WITH WEKA:

Step1: Preprocess the iris.arff dataset

Step2: Load the dataset having 5 attributes and 150 instances

Step3: Chosen the J48 classifier tool

Step4: Cross validation: 10 fold

Result:

We got the result of J48 classifier are as follows:

No of leaves: 5

Size of tree: 9

Correctly classified instances: 144 (96%)

Incorrectly classified instances: 6 (4%)

Kappa statistics: 0.94

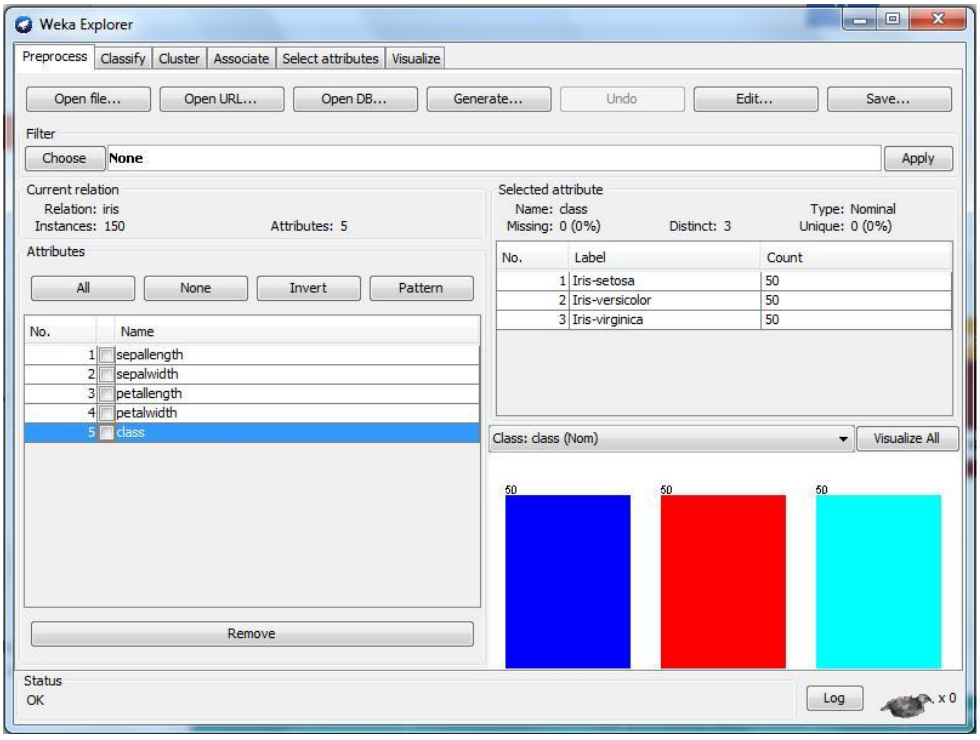

Fig. 4. Preprocessing of Dataset

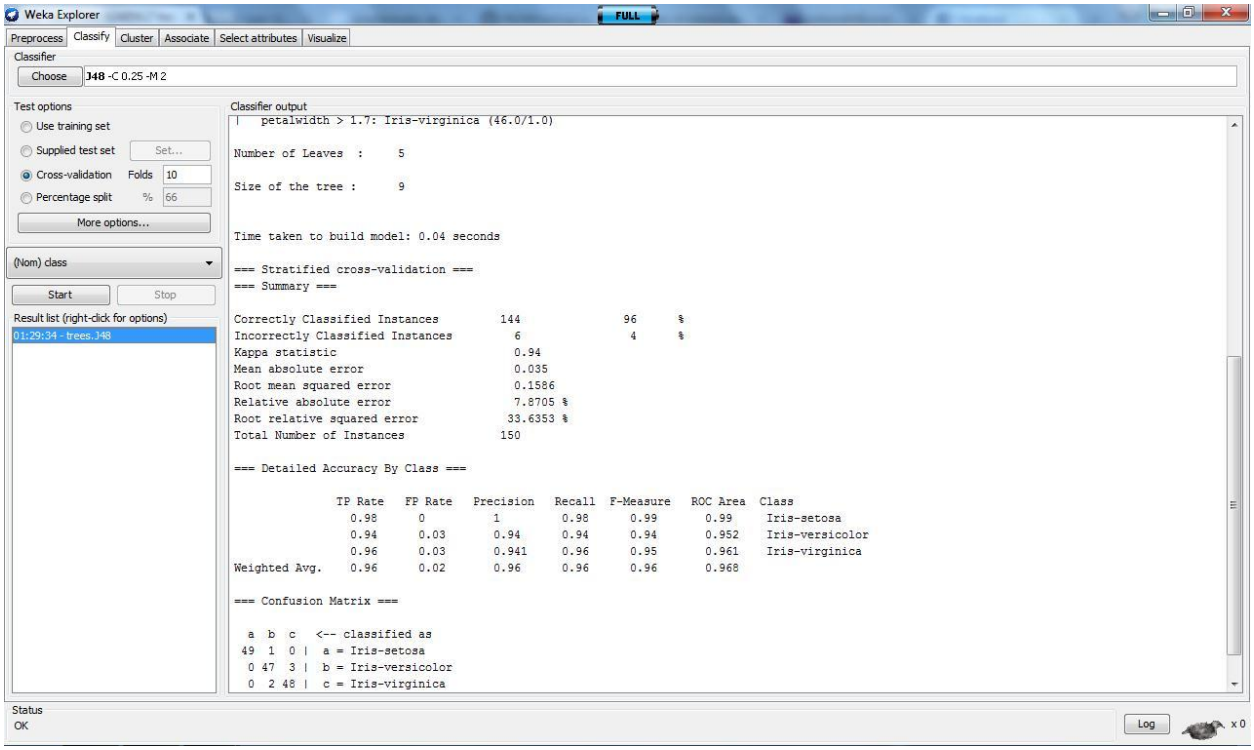

Fig. 5. Classification Result in WEKA by J48 decision tree classifier

Clustering With Weka:

Step1: Preprocess the "iris.arff" dataset from UCI data center

Step 2: Load the dataset having 5 attributes and 150 instances

Step 3: Chosen 'simple K – means clustering techniques'.

Step 4: Cluster mode: Using training dataset with percentage split 66%

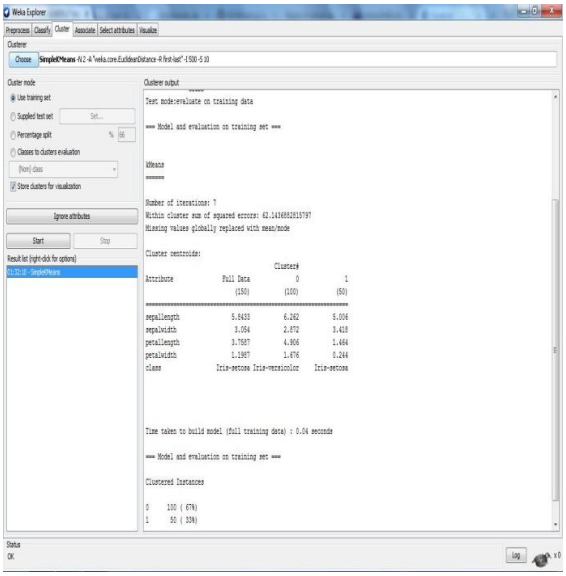

Fig. 6. Clustering through K-means algorithm in weka on IRIS dataset

Results:

We got the following results as follows:

Within cluster sum of squared error: 62.143

Cluster Instances: 0 --- 100 (67%)

1--- 50 (33%)

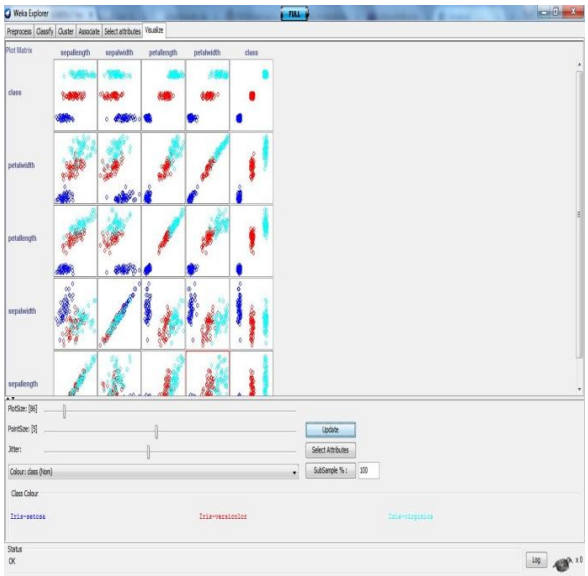

Fig. 7. Visualization of Clustering and classification result for each classes in IRIS dataset

5. CONCLUSION

This paper presents the basic classification and clustering algorithms. The comparison of decision tree, Bayesian network, k-NN, k-means, partitional clustering, hierarchical clustering and fuzzy clustering are given in this paper. The available IRIS datasets from UCI data center that researchers can utilize to carry out the research in data mining domain are listed in this paper. In the near future, interested resarcher can explore alternate ways of increasing the threshold, more accurate quality measurements, dynamic adjustment of outlier criteria and data parameters that are good indicators of how well data can be managed.

6. REFERENCES

[1] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* 11.1 (2009): 10-18.

[2] Bouckaert, Remco R. *Bayesian network classifiers in weka*. Department of Computer Science, University of Waikato, 2004.

[3] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37.

[4] Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[5] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." *IEEE transactions on systems, man, and cybernetics* 21.3 (1991): 660-674.

[6] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." *Machine learning* 29.2-3 (1997): 131-163.

[7] Peterson, Leif E. "K-nearest neighbor." *Scholarpedia* 4.2 (2009): 1883.

[8] Paterlini, Sandra, and Thiemo Krink. "Differential evolution and particle swarm optimisation in partitional clustering." *Computational Statistics & Data Analysis* 50.5 (2006): 1220-1247.

[9] Bezdek, James C., Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm." *Computers & Geosciences* 10.2 (1984): 191-203.

[10] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *International Journal of Computer Science and Applications* 6.2 (2013): 256-261.

[11] UCIDatacenter:*https://archive.ics.**uci**.edu/ml/**datasets**.html*