A COMPARISON OF MACHINE LEARNING TECHNIQUES VIA COMPUTATIONAL METHODS

Patrick Ozoh

Osun State University - Nigeria, Department of Information and Information Technology

Corresponding author: Patrik Ozoh, patrick.ozoh@uniosun.edu.ng

ABSTRACT: The process of machine learning has been useful in finding solution to issues of getting important, accurate and meaningful information. This paper provides machines with the abilities of collecting data using systems like humans and processing the data using machine learning techniques for predictions and arriving at accurate decisions the same level as humans. This paper uses two types of machine learning techniques. They techniques considered in this research are Navies Bayes and K-means clustering techniques. The confusion matrix was introduced to test the performance of the machine learning algorithms. At the end of this research, a more accurate and efficient technique would be obtained for providing insights, making reliable predictions, and for accurate decision making process.

KEYWORDS: Human senses, computational methods, predictions, classification methods, confusion matrix

1. INTRODUCTION

Machine learning expanded from traditional statistics. This has been made possible by big corporate organizations. Machine learning has been growing all through recent years. By using their procedure large amounts of data have been measured and collected. This makes it possible to utilize computational techniques to develop important models from such data [1]. There are plenty open-source implementations of machine learning techniques that can be applied using application programming interface (API) calls. For examples Weka, Orange, and RapidMiner. The results obtained can be applied to visual tools. Examples are Tableau and Spotfire. These are used for generating dashboards. Computer security can get started as a theoretically made set of events. Arranging and organizing a set of events starts with finding solutions to security issues and wants to perceive different area of computer security. Machine learning covers various rules and conditions for techniques that are applied to take out important models from collected data. These models are applied to different mining activities [2].

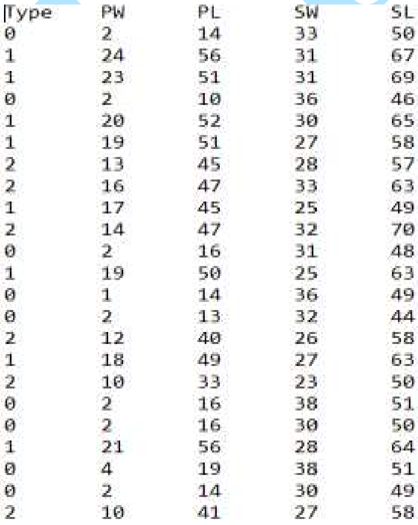
Machine learning is a branch of Artificial Intelligence (AI) and grew from pattern recognition, applied to investigate data models. It is used to model computer programs using available data. Models are used to solve complex problems. They improve the efficiency in their usage pattern.
These techniques make it easier for computers to develop models from acquired data [3]. Probability approximately correct (PAC) learning is a tool for analyzing machine learning. In applying the tool, the learner chooses a function, known as the hypothesis from a group of of feasible events. The aim is, with high probability, the chosen event will have low error. The model was enlarged to cover error and noise [4]. A major process of improving the PAC technique is the use of computational techniques to machine learning. Most importantly, the learner is expected to seek for efficient models, and the learner must develop a procedure for developing the technique. [5].

Systematic organization of information makes it easier and faster for storage, searching, and retrieval of necessary information. Text classification is a necessary tool for organizing information into groups. This is necessary because it assists organizations to improve on their manual procedures. The tool has many usage. Examples include automated indexing of articles, spam filtering, automated essay grading, classification of news articles, etc.
[6]. Information retrieval is seeking documents with solution to certain enquiries. Statistical tools can be applied to obtain stated objectives. Natural language processing can be applied to obtain better knowledge of natural language. This assists to develop classification results. Machine learning is used to design and improve techniques that allows computers to learn so as to improve system efficiency [7].

[8] applies computational techniques to analyse big data, and to analyze the behavioral pattern of such data. The paper evaluates estimated data by carrying out a simulation to accurately model actual data and at what period of time. The simulation, based on real data, is developed to estimate actual data. explores the application of effective. [9] explores machine learning to overcome challenges associated with data analysis and demonstrates how machine learning techniques have contributed and are contributing to research in machine learning. [10] focuses on developing a model that would automatically classify a comment as either toxic or non-toxic using logistic regression. The paper develops a multi-headed model to detect different types of toxicity. For example hostile words, swear words, offensive words, and racial prejudice. [11] undertakes a review of machine learning with the aim of identifying a reliable and accurate technique for modeling data. In order to identify an appropriate machine learning technique, it is necessary to carry out a comparative study of commonly used machine learning techniques.

2. DATA SET

In this research paper, the Fisher Iris data set was utilized to analyze the two machine learning techniques in this research paper. The Fisher's Iris data set is a multivariate data set introduced by [12], as an example of discriminant analysis. The Fisher Iris data set contains approximately 150 instances. The 4 features of the Fisher Iris data set are:
PL - petal length
PW - petal width
SL - sepal length
SW - sepal width
The Fisher Iris data set is given as follows:



Type	PW	PL	SW	SL
0	2	14	33	50
1	24	56	31	67
1	23	51	31	69
0	2	10	36	46
1	20	52	30	65
1	19	51	27	58
2	13	45	28	57
2	16	47	33	63
1	17	45	25	49
2	14	47	32	70
0	2	16	31	48
1	19	50	25	63
0	1	14	36	49
0	2	13	32	44
2	12	40	26	58
1	18	49	27	63
2	10	33	23	50
0	2	16	38	51
0	2	16	30	50
1	21	56	28	64
0	4	19	38	51
0	2	14	30	49
2	10	41	27	58

Fig.1 Fisher Iris data

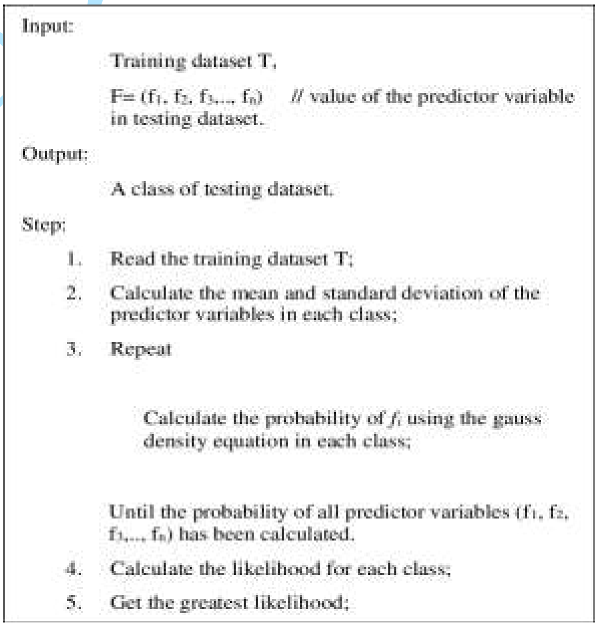
3. DESCRIPTION OF TECHNIQUES

This research paper utilizes Naïve Bayes and K-means clustering techniques as machine learning techniques to analyse data. It also utilizes comfusion matrix as a performance metrics to text the machine learning techniques. The models are described as follows:

3.1. Naive Bayes technique

Naive Bayes techniqueis a classification technique which depends on the Bayes' theorem. There exists the condition of independence between predictors. In as much as these features depend on each other, a Naive Bayes classifier contains all these properties to be independent. It is a method for developing classifiers: models for establishing class labels to problem instances, constituted as vectors of feature attributes, where the class characters are drawn from some definable set.

Naïve Bayes technique is usually applied to clustering and classification. The basic system of Naïve Bayes technique relies on the conditional probability. It establishes trees based on their probability of occurrence. These trees are as well defined as Bayesian network. The Naïve Bayes algorithm is given as follows:

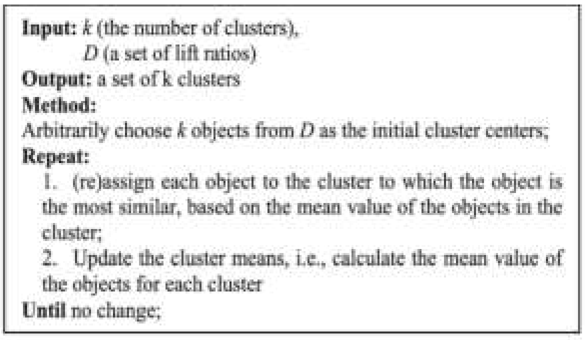


```
Input:
        Training dataset T,
        F= (f1, f2, f3,..., fn)   // value of the predictor variable
        in testing dataset.
Output:
        A class of testing dataset.
Step:
    1.  Read the training dataset T;
    2.  Calculate the mean and standard deviation of the
        predictor variables in each class;
    3.  Repeat
            Calculate the probability of fi using the gauss
            density equation in each class;
        Until the probability of all predictor variables (f1, f2,
        f3,..., fn) has been calculated.
    4.  Calculate the likelihood for each class;
    5.  Get the greatest likelihood;
```

Fig. 2 Naïve Bayes algorithm

3.2. K-means clustering technique

K-means clustering technique is defined as a set of unsupervised learning method that when it starts operating, it produces groups instinctively. The items which has close affinities are collected in the same group. This method is defined as k-means because it produces k distinct groups. The mean of the estimates in a particular group is the center of that group. K-means clustering technique is a technique of the process of constraining an input from a large set of values, collected from from signal processing, widely used for classification in data mining. K-means clustering addresses dividing n observations into k groups in which each observation share same affinity with each group. The closest mean, perform as a prototype of the group. The set of rules governing K-means clustering technique is given as follows:



```
Input: k (the number of clusters),
       D (a set of lift ratios)
Output: a set of k clusters
Method:
Arbitrarily choose k objects from D as the initial cluster centers;
Repeat:
  1. (re)assign each object to the cluster to which the object is
     the most similar, based on the mean value of the objects in the
     cluster;
  2. Update the cluster means, i.e., calculate the mean value of
     the objects for each cluster
Until no change;
```

Fig. 3 K-means clustering technique

3.3. The confusion matrix

The confusion matrix has been applied to calculate the extent of introducing and conducting the set of rules governing the method. It was applied to show the difference between the K means clustering and Naïve Bayes techniques. A confusion matrix can also be defined as the error matrix. It is a defined table structure that visualizes the system performance of the set of rules governing the method. Individual row of the matrix denotes variables in a predicted group while each column denotes variables in a group. It is a unique type of contingency table, containing two dimensions, actual and predicted, and also similar types of groups in both dimensions of the contingency table.

4. IMPLEMENTATION AND RESULTS

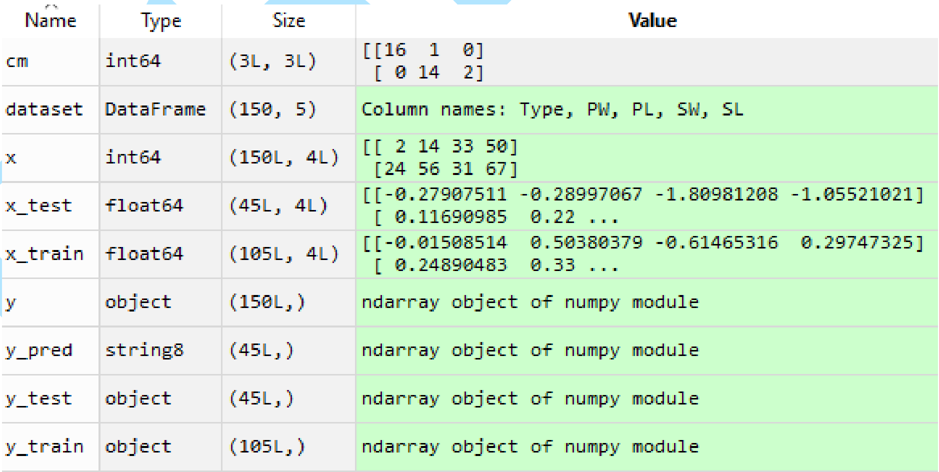
This section presents the implementation of the machine learning techniques via computational method. Implementation results for K-means clustering and Naives Bayes techniques are given. The results of the performance of the K means clustering and the Naïve Bayes techniques using confusion matrix are also given.

4.1. Results for K-means clustering technique

After the analysis of data sets using K-means clustering algorithm, the following results are obtained:

4.2. Results for Naives Bayes technique

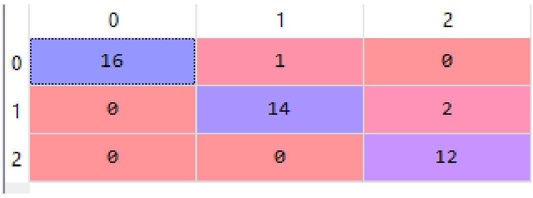
After the analysis of datasets using Naives Bayes technique algorithm, the following results are obtained:



Name	Type	Size	Value
cm	int64	(3L, 3L)	[[16 1 0] [0 14 2]
dataset	DataFrame	(150, 5)	Column names: Type, PW, PL, SW, SL
x	int64	(150L, 4L)	[[2 14 33 50] [24 56 31 67]
x_test	float64	(45L, 4L)	[[-0.27907511 -0.28997067 -1.80981208 -1.05521021] [0.11690985 0.22 ...
x_train	float64	(105L, 4L)	[[-0.01508514 0.50380379 -0.61465316 0.29747325] [0.24890483 0.33 ...
y	object	(150L,)	ndarray object of numpy module
y_pred	string8	(45L,)	ndarray object of numpy module
y_test	object	(45L,)	ndarray object of numpy module
y_train	object	(105L,)	ndarray object of numpy module

4.3. Performance test using confusion matrix

The result from the implementation of the confusion matrix was used to show the performance of the K means clustering and the Naïve Bayes techniques. The results from implementing the confusion matrix on Naïve Bayes technique is given as follows:



	0	1	2
0	16	1	0
1	0	14	2
2	0	0	12

Fig. 4 The results on Naïve Bayes technique

The first row (0,1,2) shows the predicted types while the first column (0,1,2) shows the real types. The values in blue were the rightly predicted valve.

(0,0): shows that 16 values was correctly predicted to be type 0

(1,1): shows that 14 valves was correctly predicted to be type 1

(2,2): shows that 12 values was correctly predicted to be type 2

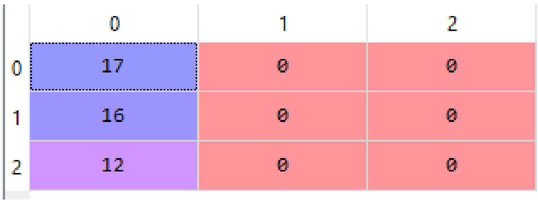
(0,1): shows that 1 value was wrongly predicted to be type 1

(1,2): shows that 2 valve was wrongly predicted to be type 2

(1,0), (2,0), (2,1), (0,2): shows that there is no valve predicted.

The Naïve Bayes shows that three valves are predicted rightly and two are wrongly predicted.

The results from implementing the confusion matrix on K means clustering technique is given as follows:



	0	1	2
0	17	0	0
1	16	0	0
2	12	0	0

Fig. 5 The results on K means clustering technique

The first row (0,1,2) shows the predicted types while the first column (0,1,2) shows the real types. The values in blue were the rightly predicted valve.

(0,0): shows that 17 values were correctly predicted to be type 0

(1,0): shows that 16 valves were correctly predicted to be type 1

(2,0): shows that 12 values were correctly predicted to be type 0

The K means clustering shows that three valves are predicted rightly.

5. SUMMARY AND CONCLUSIONS

This research provides machines with the abilities of collecting data using systems identical to the humans and then processing collected data via machine learning methods for predictive purposes and inferring decisions the same level as humans. In evaluating the objectives, the study employed two machine learning techniques. The confusion matrix was introduced to test the performance of the machine learning algorithms.

From the analysis carried out, Naives Bayes technique was able to predict more accurate values than K means clustering technique. This study has added to already numerous literature on machine learning techniques. The study has contributed to show the differences between K means clustering technique and Naives Bayes algorithm using confusion matrix.

REFERENCES

[1] **R. Owen** - *Polynomial Approximations to the Cumulative Fisher Distribution*, Educational and Psychological Measurement. vol. 32, no. 1, 313-319, 1972.

[2] **T. Hastie, R. Tibshirani, J. Friedman** - *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Springer, New York, 2nd edition, 2009.

[3] **A. Krizhevsky, L. Sutskever, G. Hinton** - Advanced Neural Inference Process, Journal Systems, vol. 25: 1097–1105, 2015.

[4] **D. Levy** - *Computer Games I*, Springer, US, 2011.

[5] **K. Yu, W. Lam** - *A new on-line learning algorithm for adaptive text filtering*, Proceedings of the 7th International Conference on Information and Knowledge Management, Betheseda, US: 156–160, 1998.

[6] **M. Maron** - *Automatic indexing: an experimental inquiry,* Journal of the Association for Computing Machinery, vol. 8, no. 3: 404–417, 1961.

[7] **P. Ozoh, M. Olayiwola, I. Ogundoyin** - *Analysis of Computational Techniques to Analyze Big Data*, Communications in Applied Sciences, vol. 8, no. 1: 1-18, 2020.

[8] **D. Delen** - *A comparative analysis of machine learning techniques for student retention management - Decision Support Systems,* vol. 49, no. 4: 498-506, 2010.

[9] **P. Vidyullatha, S. Padhy, J. Priya, K. Srija, S. Koppisetti** - *Identification and Classification of Toxic Comment Using Machine Learning Methods*, Turkish Journal of Computer and Mathematics Education, vol 12, no 9: 70-74, 2021.

[10] **P. Ozoh, M. Olayiwola, A. Adigun** - *An In Depth Study of Typical Machine Learning Methods via Computational Techniques*, Annals. Computer Science Series, vol. 16, no. 2: 77-81. 2018.

[11] **L. Valiant** - *A theory of the learnable*, Communications of the ACM, 1984.

[12] **Y. Yang** - *An evaluation of statistical approaches to text categorization,* Journal of Information Retrieval 1, vol. 1, no.2: 69–90, 1999.