iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37773

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

A Comparative study of C4.5 and Naïve Bayes Algorithms

Raksha Shukla[1], Dr. Sanjay Kumar[2],
[2]*Ph.D. Scholar,* [1]*Professor, Computer Science, Kalinga University, Raipur (C.G.), India*

Abstract: ***In this fast and growing age large amount of data is generated and save or store somewhere daily. These data are included many information. They may be financial data, scientific, medical science related data, engineering data and many other types of data. Analyzing such data is an important need. Data Mining is a technique which providing tools to discover knowledge from data. It includes so many techniques for KDD such as classification, clustering, regression etc.***
In this paper a comparative study of C4.5 and Naïve Bayes Data mining classification techniques has been done. The experiment is based on weather data. That includes parameters like temperature, humidity, wind speed etc. Weka tool is used here for the experiment.
Keywords: Data Mining, Classification, Naïve Bayes, C4.5.

I. INTRODUCTION

Data Mining should have been more appropriately named "Knowledge mining from data" as per Han and Kamber [1]. Many other terms have a similar mining to data mining for example- knowledge mining from data, knowledge extraction, data pattern analysis, data archaeology and data dredging. Many people treat data mining as a synonym knowledge discovery from data or KDD. This has following steps-

A. Data cleaning (to remove noise and inconsistent data)
B. Data Integration(multiple data sources may be combined)
C. Data Selection(where data relevant to the analysis task are retrieved from the database)
D. Data Transformation(where data are transformed and consolidated in appropriate forms)
E. Data Mining(an essential process where intelligent methods are applied to extract data patterns)
F. Pattern Evaluation(to identify the truly interesting pattern representing knowledge based on interesting pattern)
G. Knowledge Presentation(visualization and knowledge representation techniques are used to present mined knowledge to users)

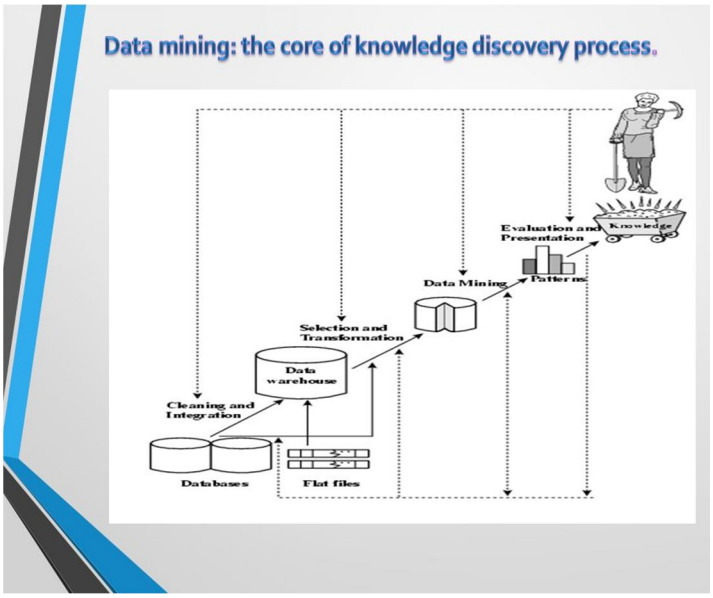


Figure-1: KDD Process

Table -I: Review of Literature

Year	Authors	Techniques/Algorithms	Dataset/Attributes	Research Work
2016	Dr. B.Srinivasan and K.Pavya	C4.5, ID3, KNN, SVM, ANN, Naïve Bayes		Comparative Study and their pros and cons
2017	Mrs. Nalini Jagtap, Mrs. P.P. Shevatekar,	Decision tree, NN, SVM, generic algorithms and fuzzy logic		Faced some pros and cons with growing volume of data using algorithms
2018	Muhammad Alghobiri	Naïve Bayes, C4.5, SVM	Diverse Datasets	SVM performs best as comparison of others
2019	Alpa Makvana, Devangi Kotak	SVM, NN	14 attributes	Both techniques are well
Jan 2021	Madhu Kahar, Manisha Kahar	SVM, KNN, ANN, Naïve Bayes, Clustering, Prediction, Association rule		Classification technique provides high accuracy
Mar 2021	Luis Chaves and Goncalo Marques	Naïve Bayes, NN, Adaboost, KNN, Random Forest, SVM	Diabetes Dataset(520)/17	Neural Network should be used for diabetes prediction

II. LITERATURE REVIEW

There are so many research works have been done for the comparisons of two or more classification techniques. Datasets are applied on those algorithms and comparison made on the basis of their performances. Data may collect from data Repository or related institutions. Some of the research works are discussed in above table-

III. METHODOLOGY

A. Classification

Classification is a main domain of data mining as per Muhammad Alghobiri [4] technique which maps data into predefined groups or classes. It is supervised learning because classes are defined before examining of data.

Classification is used for different purposes like machine learning, pattern recognition, network security, medical science **L**uis Chaves and Goncalo Marques [7] etc. There are many classification techniques decision tree, Naïve Bayes, KNN, SVM etc.

1) C4.5 Algorithm: C4.5 is an algorithm is used to generate decision tree using information gain in the same way as ID3 algorithm developed by Ross Quinlan as a successor of ID3 algorithm. It is used for great volume of data so that it will be helpful to generate best classification and consists for large decision trees. C4.5 algorithm can handle missing attribute value and handles attributes with different cost. This algorithm is easy to understand as compare to ID3 algorithm.

a) Steps

- All the attributes in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class detected. Again, C4.5 creates a decision node higher up the tree using the expected value.
- For each attribute a, find the normalized information gain ratio from splitting on a.
- Let a_best be the attribute with the highest normalized information gain as per Dr.B.Srinivasan and K.Pavya[2].
- Create a decision node that splits on a_best.
- Recur on the sub lists obtained by splitting on a_best, and add those nodes as children of node.

b) Advantages of C4.5 algorithm

- It handles both continuous and discrete items.
- It handles training data with missing attributes.
- Handling attributes with different costs.
- Pruning trees after creation of decision tree.

2) *Naïve Bayes:* Naive Bayes Classifier is a simple technique for classifier that depends on Bayes' theorem with independent assumptions. It provides data structure and facilities common to Bayes network learning algorithms. An advantage of Naive Bayes Classifier is that it only requires a small number of training data to estimate the attributes for classification [1].

a) Bayes' Theorem: Probability (B given A) = (Probability (A and B) / Probability (A))

Advantages of Naive Bayes Classifier:

- Easy handle of large amount of data.
- Handles real and discrete data.

B. Weka(3.8.3)

Waikato Environment for Knowledge Analysis (Weka), as given in Wikipedia[8] is a data mining/machine learning tool developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License, Weka is a bird found only in New Zealand. This is a collection of machine learning algorithms for data mining tasks. The algorithm can either be applied directly to a dataset or called from Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rools and visualization. It is also well suited for developing new machine learning schemes. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

C. Dataset

Weather data is used for this research paper. The dataset has been taken from UCI Machine learning repository. It is 1(2012) year weather data for Washington City so there is 366 instances. It consists three attributes precipitation, temperature and wind.

IV. COMPARATIVE ANALYSIS

In this section result of both classification algorithms shown on the Table. A chart for the given result is also put here. Based on the find outs decision can be made.

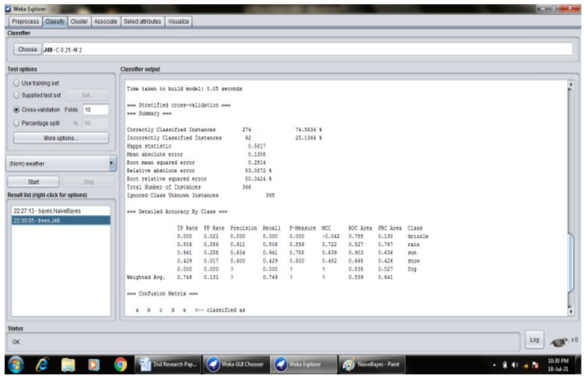


Figure-2: Classification by J48

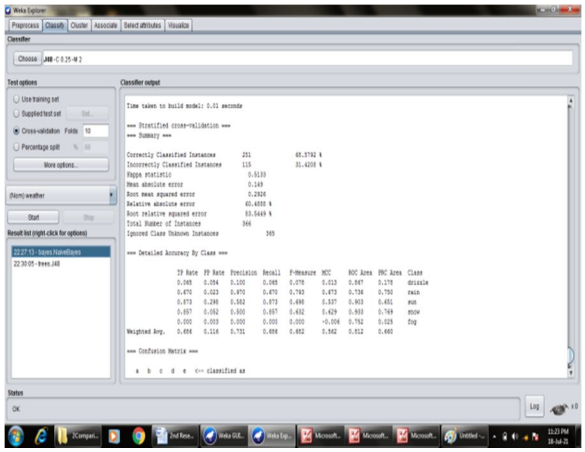
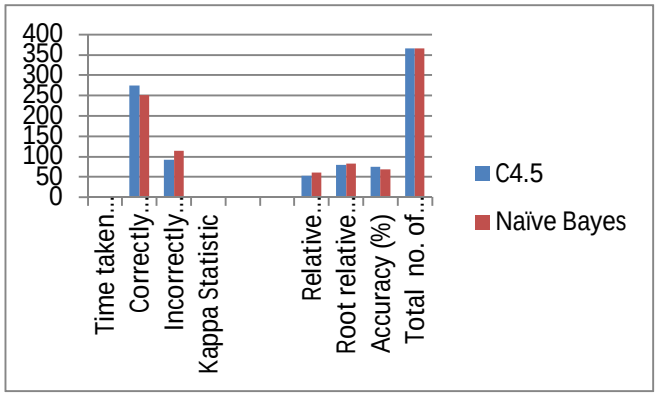


Figure-3: Classification by Naïve Bayes

TABLE-II: Comparison OF C4.5 And NAÏVE Bayes

Sq. No.	Performance	C4.5	Naïve Bayes
1.	Time taken to build model (seconds)	0.05	0.01
2.	Correctly Classified Instances	274	251
3.	Incorrectly Classified Instances	92	115
4.	Kappa Statistic	0.5817	0.5513
5.	Mean absolute error	0.1308	0.149
6.	Root Mean square error	0.2814	0.2926
7.	Relative absolute error (%)	53.0872	60.4888
8.	Root relative squared error (%)	80.3424	83.5449
9.	Accuracy (%)	74.8634	68.5792
10.	Total no. of instances	366	366

Chart-1: Comparison Chart

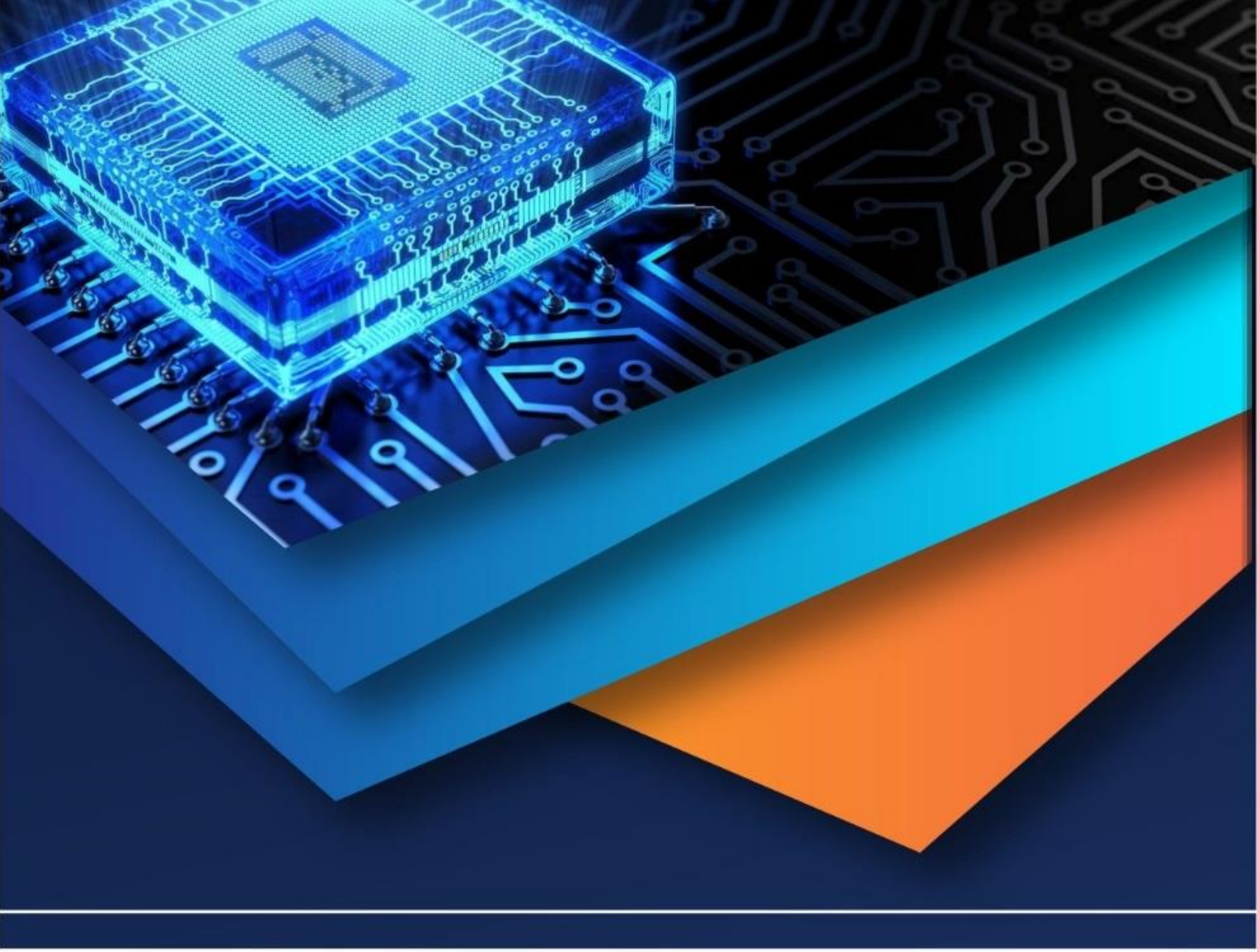


V. CONCLUSION

From the above result and analysis the conclusion is that C4.5 is better than Naïve Bayes algorithm for the prediction of weather dataset because C4.5 Model accuracy is greater than Naïve Bayes and model build up time of C4.5 is also less then Naïve Bayes as well.

REFERENCES

[1] J iawai Han and Micheline Kamber Data Mining: Concepts and Techniques, 3rd edition.
[2] Dr.B.Srinivasan and K.Pavya "A Comparative Study on Classification Algorithms in Data Mining" IJISET, Vol. 3 Issue3, March 2016.
[3] Mrs. Nalini Jagtap, Mrs. P.P. Shevatekar, Mr. Naresh Kumar Mustary . "A Comparative Study of Classification Tehniques in Data Mining Algorithms". IJMTER 2017.
[4] Muhammad Alghobiri "A Comparative Analysis of Classification Algorithms on Diverse Dataset" Engineering, Techonology and Applied Science Research. Vol. 8,No. 2, 2018,270-275.
[5] Alpa Makvana, Devangi Kotak "Comparative Analysis of Data Mining Classification Techniques for Cardiovascular Disease Prediction" IJERT Vol 8 Issue 11, Nov 2019.
[6] Madhu Kahar, Manisha Kahar "A Comparative Analysis of Data Mining Algorithms and Techniques" Vol 3 Issue 01, Jan 2021.
[7] Luis Chaves and Goncalo Marques "Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study" applied sciences MDPI Appi Scie.2021.
[8] Wikipedia.
[9] Y A Gerhana et al 2019 J. Phys.: Conf. Ser. 1280 022022 "Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period" IOP publishing 2019.
[10] Aakanksha Jain et al 2021 J. Phys.: Conf. Ser. 1854 012046 "Comparative study of Jrip j48 and naive bayes algorithm in Flower specie prediction" IOP publishing 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY