International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

COMPARATIVE ANALYSIS OF NAÏVE BAYES AND HILL CLIMBER SEARCH ALGORITHMS IN DATA MINING USING WEKA TOOL

Jaspreet Singh Phul[1], Sonia Vatta[2]

[1]*Research Scholar, Bahra University, HP, India*

[2] *AP, Bahra University, HP, (India)*

ABSTRACT

Data Mining is the process of extracting useful information from database after summarizing it. In medical area, data mining plays important role to discover new patterns to provide useful and meaning information. Now a days, Data Mining techniques and search algorithms are applied to healthcare datasets to analyze the diabetes process. The Naïve Bayes algorithm is basically used for prediction and exploratory modeling and to discover relationships between input and predictable columns. The hill Climber, the diabetes dataset with a total sample records 768 and 9 attributes (8 for input and 1 for output) will be used to test. The aim of this paper is to compare search algorithms i.e., Naïve Bayes and Hill Climber and evaluate results by applying on small and large dataset and find which is best and second best.

Keywords: Data Mining, Hill Climber, HealthCare, Naive Bayes, Search Algorithm.

I. INTRODUCTION

Data Mining is a process of extracting useful information and transforms into structure form for further use and also find patterns in dataset. Classification techniques in healthcare can be applied for diagnosis purposes based on some criteria. Classification is also applied to a wide range of application areas such as weather prediction, education, customer segmentation in banking etc.Many classification techniques such as decision tree, J48craft are used to predict disease. The main focus of this paper is to compare search algorithms. After comparison, evaluation process is performed based on small and large dataset to check either it give different results or same results and which is best. To fulfill this process, discretized data is assumed. Each input variable discretized into three section i.e. “low”,”medium”,”high” by using search algorithms.

II. METHODOLOGY

The two search algorithms are used to find the best algorithm for diabetes dataset on the 10 fold Cross-validation and percentage split. The comparative analysis is given below Fig.1:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

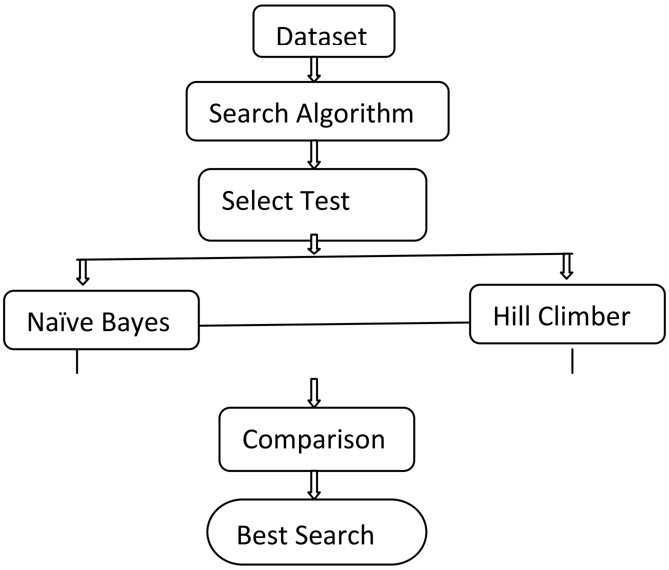


Fig.1: Flow Chart for Analysis.

Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Disease from UCI dataset used for data mining classification. The dataset contains 768 Instances (record samples), each having eight attributes. Consider no missing values. Table 1 shows attribute description.

Table 1: Attribute Description.

Sr. No.	Attribute	Relabeled values
1.	Number of times pregnant	Preg
2.	Plasma glucose concentration	Plas
3.	Diastolic blood pressure (mm Hg)	Pres
4.	Triceps skin fold thickness (mm)	Skin
5.	2-Hour serum insulin	Insu
6.	Body mass index (kg/m2)	Mass
7.	Diabetes pedigree function	Pedi
8.	Age (years)	Age
9.	Class Variable	(0 or 1) Class-Not applicable

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

III. PRE PROCESSING DATA

The first step in data mining process is to process the data. For this, load diabetes data from data folder located in Weka dataset.

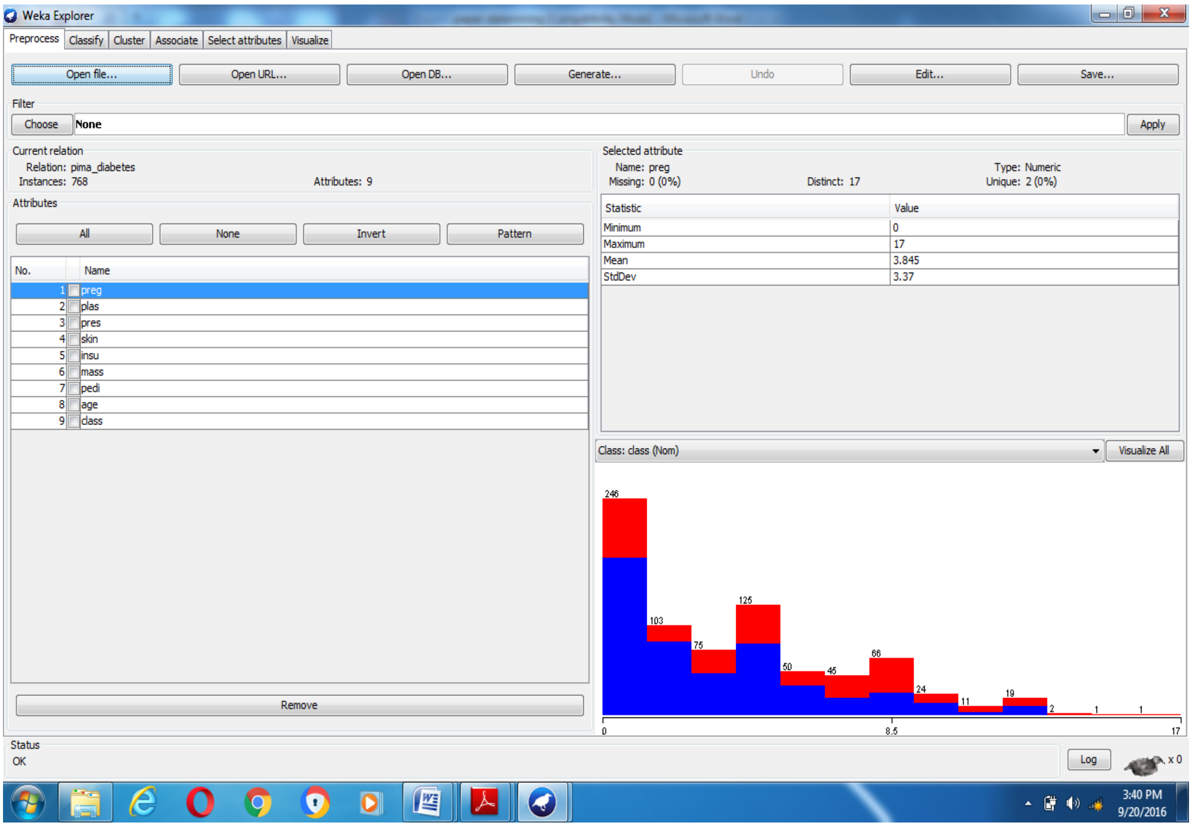


Fig. 2: Diabetes Datasets Open in Weka

After selecting dataset, next step is to choose filters to transform the input data. Now select discredited attribute of unsupervised learning and allow Useequalfrequency property to be true. For this, class variable is not necessary to consider.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

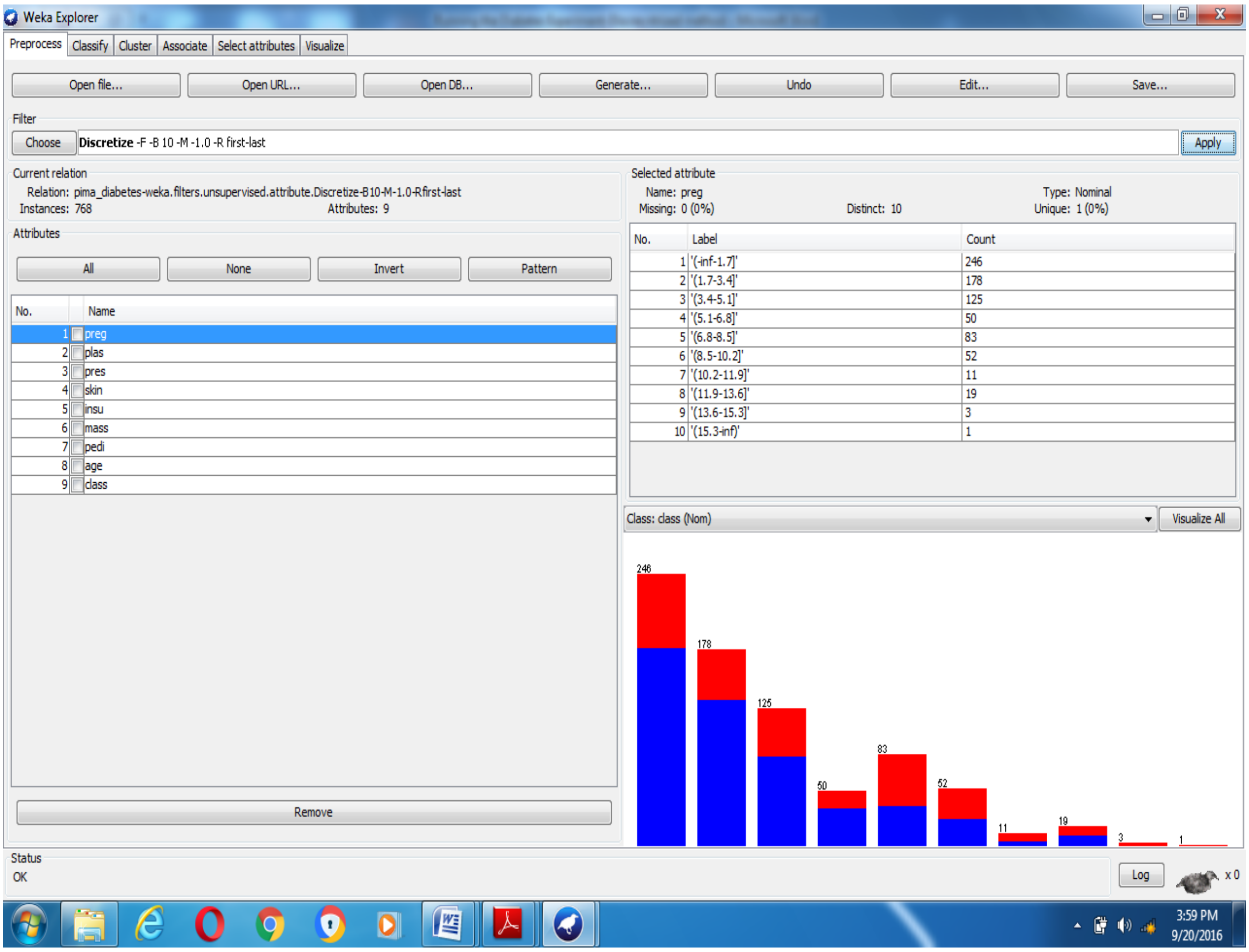
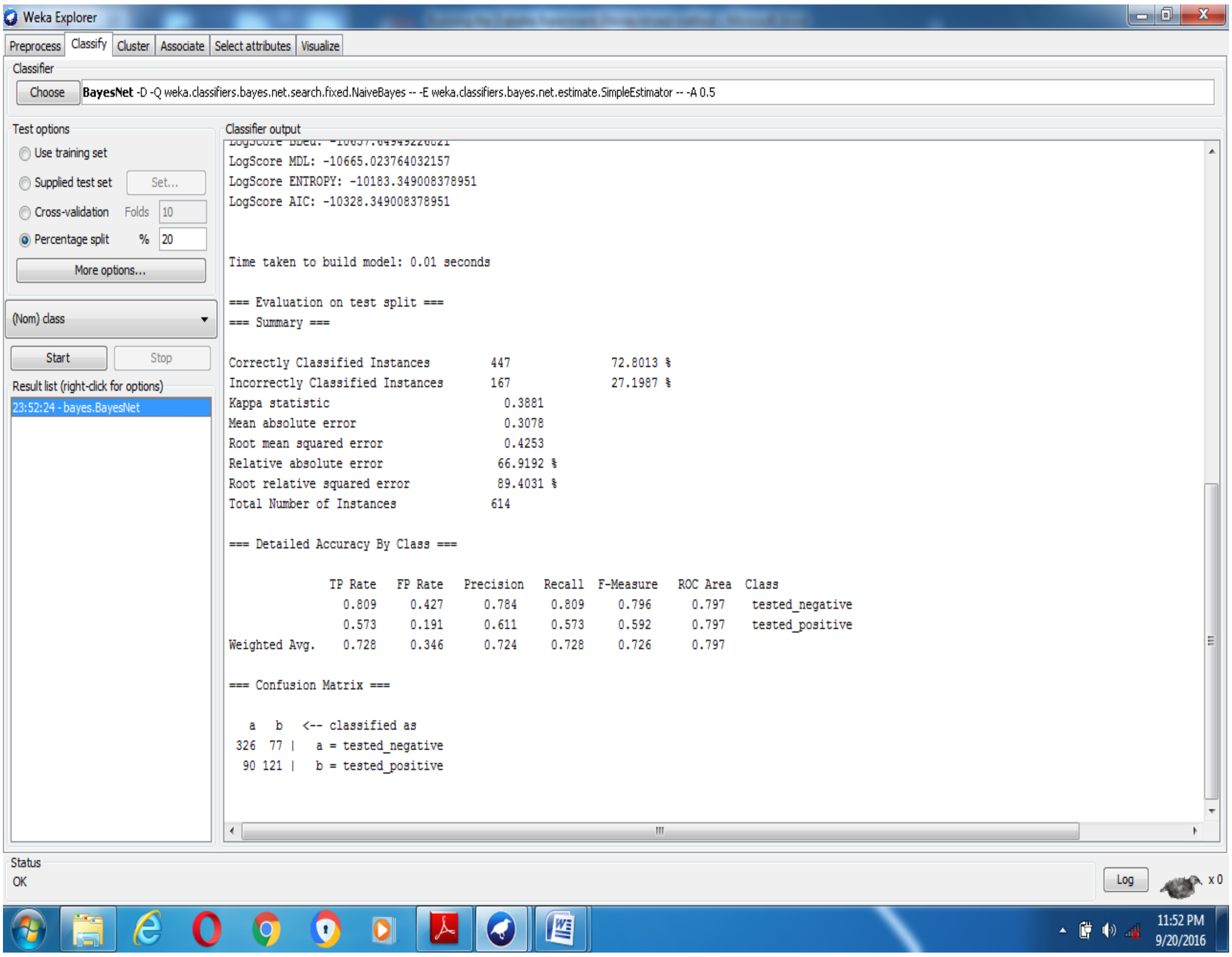


Fig. 3: Discretized Process

IV. TESTING PROCESS FOR NAIVE BAYES ALGORITHM

The Naïve Bayes classifier is based on the Bayes rule of conditional probability. It makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other. From "Weka window", select percentage split option to evaluate the quality of the model from "test options" section. First consider for small dataset i.e., two values for test and divide values in terms of percentage 20% and 70%.The result for 20% is obtained as follows:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346



The confusion matrix is:

Confusion Matrix ===

a b <-- classified as

326 77 | a = tested_negative

90 121 | b = tested_positive

From confusion matrix where “a” denotes the patients having no diabetes. Hence, there are a total of 326+77=403 patients without diabetes and b denotes the patients having diabetes. Therefore, there are 90+121=211 patients with diabetes.

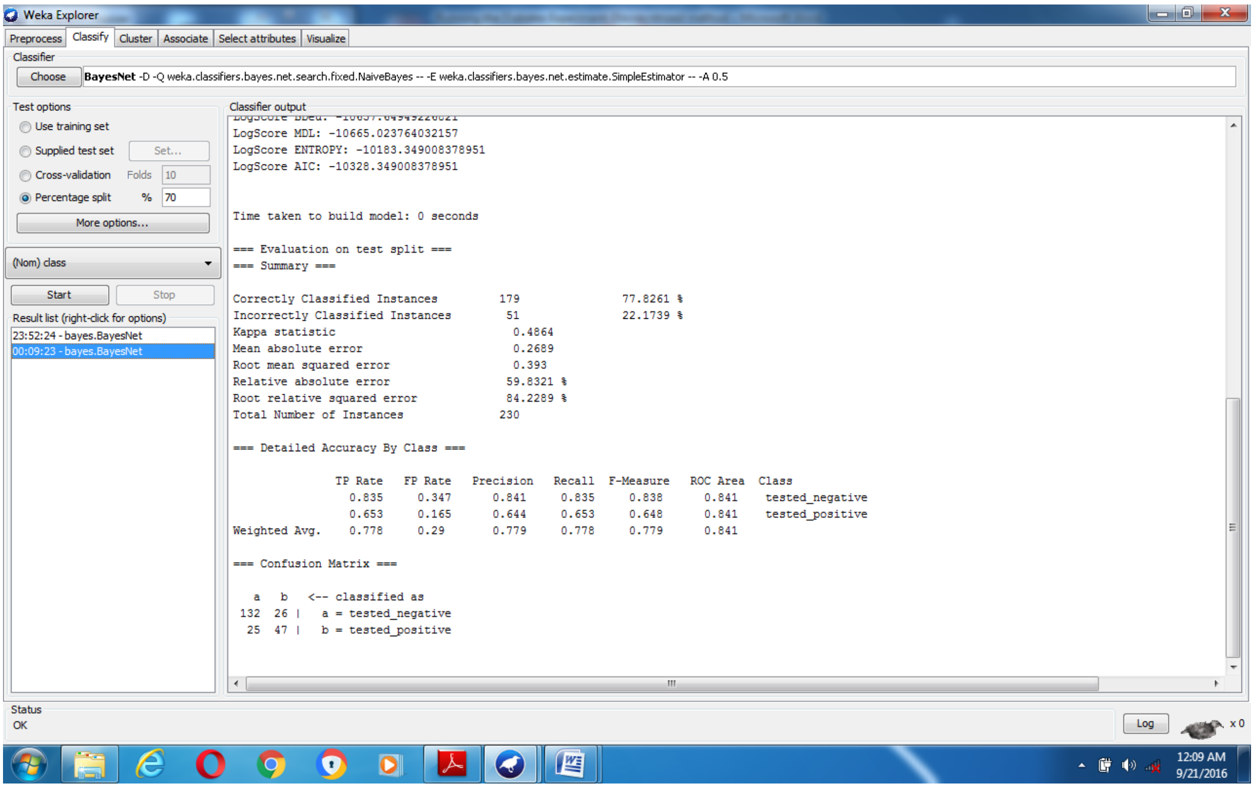
International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Sr. No.	Correctly Classified(Negative)	Incorrectly Classified(Positive)	Result
1.	326	77	tested_negative
2.	90	121	tested_positive

Table 2: Matrix corresponds for small sample.

V. TESTING PROCESS FOR NAIVE BAYES ALGORITHM (Now test for 70% (large Sample).

The result is obtained as follows:



From confusion matrix, we found that the error rate is now lower as compared to small samples selection. We have only 132 negative cases and 25 positive cases in test dataset.

Sr. No.	Correctly Classified(Negative)	Incorrectly Classified(Positive)	Result
1.	132	26	tested_negative
2.	25	47	tested_positive

Table 3: Matrix corresponds for large sample.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

I. Testing Process for Hill Climber Algorithm(For small samples (20%)

The result is obtained as follows:

Sr. No.	Correctly Classified(Negative)	Incorrectly Classified(Positive)	Result
1.	325	78	tested_negative
2.	81	130	tested_positive

Table 4: Matrix corresponds for small sample.

II. TESTING PROCESS FOR HILL CLIMBER ALGORITHM (For large samples (70%))

The result is obtained as follows:

Sr. No.	Correctly Classified(Negative)	Incorrectly Classified(Positive)	Result
1.	128	30	tested_negative
2.	26	46	tested_positive

Table 5: Matrix corresponds for large sample.

VI. CONCLUSION

Both the algorithms are applied on the diabetes dataset and the results are given in table 2, 3, 4, and 5. From the result we see time to build the model is less when using Hill Climber and correctly classified instances are more when using Hill Climber and prediction accuracy is also greater in Hill Climber than of Native Bayes .Why Native Bayes is second best? The reason is that Naïve Bayes classifier requires a very large number of records to obtain good results. Second, where a predictor category is not present in the training data, Naive Bayes assumes that a new record with that category of the predictor has zero probability. When it classifies, performance does not show significant improvement. Hence it is concluded that Hill Climber worked best for both small and large dataset as compared to Native Bayes.

REFERENCES

[1] Remco R. Bouckaert, Eibe Frank, Mark Hall Richard Kirkby, Peter Reutemann, Seewald David Scuse, WEKA Manual for Version 3-7-5, October 28, 2011.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[2] Anshul Goyal , Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", Published in International Journal of Applied Engineering Research, ISSN: 0973-4562 Vol.7 no.11 (2012).

[3] C. M. Velu and K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients", 3rd IEEE International Advance Computing Conference (IACC), 2013

[4] Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[5] http://en.wikipedia.org/data-mining.

[6] http://www.emedicinehealth.com/diabetes.

[7] Pardha Repalli, "Prediction on Diabetes Using Data mining Approach".

[8] Huan Liu, Hiroshi Motoda. Feature selection for knowledge discovery and data mining.