International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET218410

Observing Performance of Naive Bayes Classifier on Nursery Dataset

Rajni Bhalla, Amandeep
Lovely Professional University, Phagwara, Punjab, India

Article Info
Volume 9, Issue 3
Page Number : 91-95
Publication Issue :
May-June-2022
Article History
Accepted : 01 May 2022
Published: 13 May 2022

ABSTRACT

In machine learning, Naive Bayes is a popular technique that is used for classification that is based on the conditional probability of attributes belonging to a label, in which the attribute is selected by select attribute operator in rapid miner. In this paper, the split operator has used that divides the dataset into training and testing. Training is used to train the naïve Bayes and testing is used to evaluate the model. The result shows that this simple model generates a good fit for the nursing dataset. Total accuracy achieved using this method is 87.86% which is not bad.

Keywords: Naïve Bayes, Performance, Classification, Data mining

I. INTRODUCTION

In Machine learning, there are a number of methods used for classification. One of the methods is naive Bayes that belong to a family of probabilistic classifiers. Naïve Bayes is based on Bayes theorem [1]. Kernel density estimated can be coupled to achieve the highest accuracy [2] [3]. Naïve Bayes was introduced in text retrieval in the 1960s. It is a popular method and used for text categorization also. This is a classification method that is used for judging whether it belongs to one category or another. Before applying naïve Bayes, preprocessing is required to perform on the dataset [4]. Naïve Bayes has been used and implemented by a number of researchers for performing classification [5] [6]. Naïve successfully used for predicting diabetic, heart disease, and mobile phone [7][8][9]. In this paper, naïve Bayes is based on the conditional probability of the attributes belonging to one class after attribute selection done by using the select attribute operator. The experiments with the dataset are obtained from UCI repositories on which naïve Bayes is applied to check the performance of the model.

II. METHODS AND MATERIAL

2.1. Dataset

This dataset has been taken from UCI repositories. https://archive.ics.uci.edu/ml/datasets/nursery.
It was resulting from a hierarchical decision model. The main motive to develop this dataset is to grade submission for nursery school. This dataset consists of 12,960 examples. There are no missing values in this dataset. There are a total of 9 attributes in this dataset. Explanations of attributes are given below:-
1. Parents: This feature has information about the parents of the kid. It consists of three values: usual, pretentious, and great_pret. It consists of three values, it will also set to polynomial in rapidminer.
2. Form: All the information related to the form filled by the applicant is given in this attribute. It has five possible values: proper, less_proper, improper, critical,

very_crit. In rapidminer, the datatype of this attribute will set to polynomial because it consists of five values.

3. Has_nur: This attribute has information on whether the nursery of the child is proper, less_proper, improper, critical, and very_crit. The data type of attribute will set to polynomial.

4. Children: This attribute consists of information about the number of children about the applicant whether it has one, two, three, or four. It consists of four values so this datatype will set to polynomial.

5. Housing: Housing standard of the applicant defined by possible values like: convenient, less_conv, and critical. These attribute values will set to polynomial.

6. Finance: The financial standing of the applicant will be defined by possible values: convenient, inconv. This attribute consists of two values so datatype will set to binomial.

7. Social: This attribute defines the social structure of the family that consists of three values: nonprob, slightly_prob, and problematic. The data type of this attribute will set to polynomial.

8. Health: This attribute defines the health picture of the family. This attribute has three values: recommended, priority, and not recommended. The data type of this attribute will also set to polynomial.

9. Rank: This is the target attribute. The role of this attribute will be set as a label. It specifies the rank of the application. This attribute consists of five values: not_recom, recommend, very_recom, priority, and spec_priority. The value for this attribute will be predicted by classification algorithms. Except for rank attribute, the role of other attributes will be set to regular.

2.2 Operators to perform classification

The first step is import dataset using read excel operator. Nursery dataset is imported using 'read excel' operator in rapid miner as shown in Figure 1[10]. When we select read excel operators, all properties related to the operator will be shown on the right-hand side. An Import configuration wizard is used for the loading dataset. Excel file will display and we can change the role and type of attributes. Rank attributes will set the role of label attributes. Rest all attributes role be set as regular.

5000 rows will be imported using a rapid miner. The select attribute is used for selecting only the has_nurs, health, and parent attributes that are selected to simplify this process.

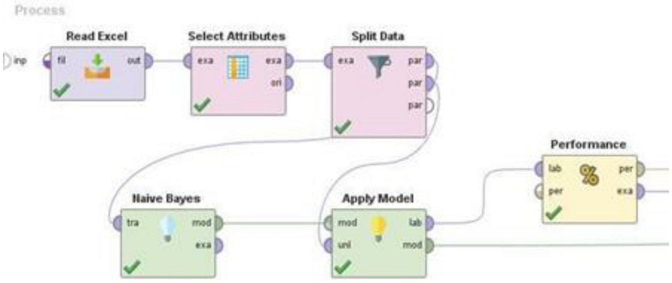
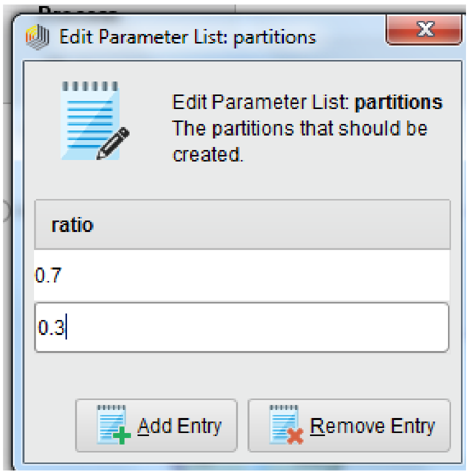


Figure 1 Naive Bayes operator

After successfully importing the dataset, the next step is to apply pre-processing steps on the dataset. We apply the split operator to divide the dataset into training and testing and then evaluate. Splitting is performed to check the performance of the operator. The split operator is used, when we want to check the accuracy of the model. It is divided into two sub processes. One is known as training and the other is known as testing. The training process is used to create a model. This partition is used to train a naïve Bayes classification model and the splitting ratio for train and test is 0.7 and 0.3 as shown in Figure 2. After creation, that model is applied to the testing dataset to check the performance of the new model.



Edit Parameter List: partitions

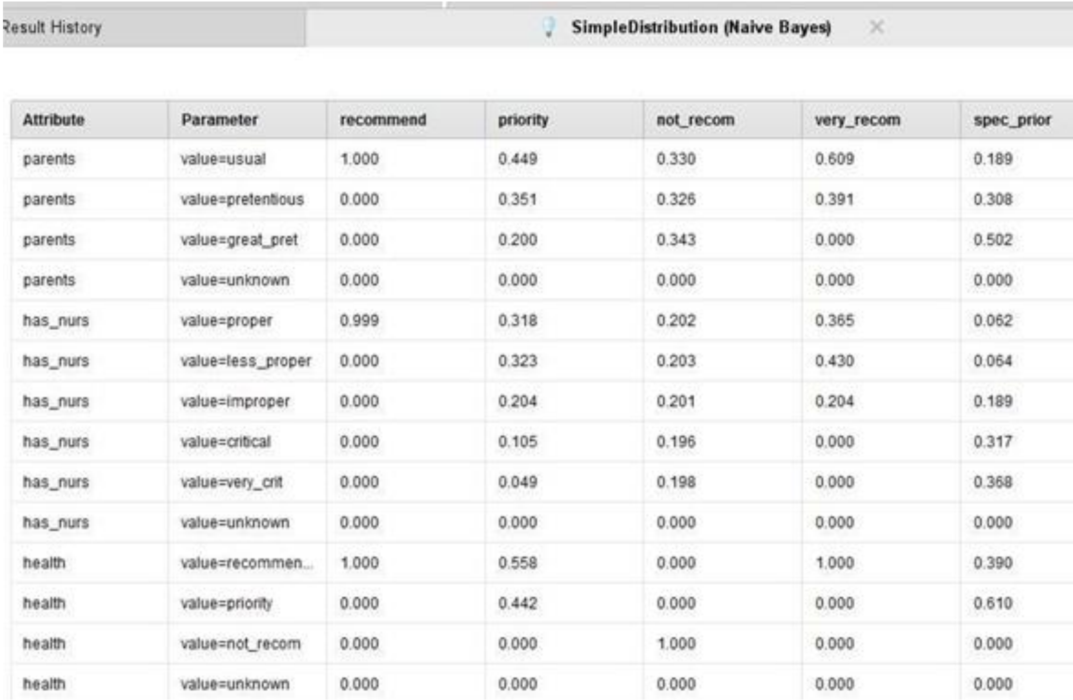
Edit Parameter List: **partitions**
The partitions that should be created.

ratio
0.7
0.3

Figure 2 Parameters of Split Data operator

III. RESULTS AND DISCUSSION

The result is shown below.



Result History

SimpleDistribution (Naive Bayes)

Attribute	Parameter	recommend	priority	not_recom	very_recom	spec_prior
parents	value=usual	1.000	0.449	0.330	0.609	0.189
parents	value=pretentious	0.000	0.351	0.326	0.391	0.308
parents	value=great_pret	0.000	0.200	0.343	0.000	0.502
parents	value=unknown	0.000	0.000	0.000	0.000	0.000
has_nurs	value=proper	0.999	0.318	0.202	0.365	0.062
has_nurs	value=less_proper	0.000	0.323	0.203	0.430	0.064
has_nurs	value=improper	0.000	0.204	0.201	0.204	0.189
has_nurs	value=critical	0.000	0.105	0.196	0.000	0.317
has_nurs	value=very_crit	0.000	0.049	0.198	0.000	0.368
has_nurs	value=unknown	0.000	0.000	0.000	0.000	0.000
health	value=recommen...	1.000	0.558	0.000	1.000	0.390
health	value=priority	0.000	0.442	0.000	0.000	0.610
health	value=not_recom	0.000	0.000	1.000	0.000	0.000
health	value=unknown	0.000	0.000	0.000	0.000	0.000

Figure 3 Distribution table

From Figure3, we can see how confidence for the first example is calculated. The rows where “parents=usual”, “has_nurs=proper”, health=” recommended”. All these rows are used for calculations for confidences in this example.

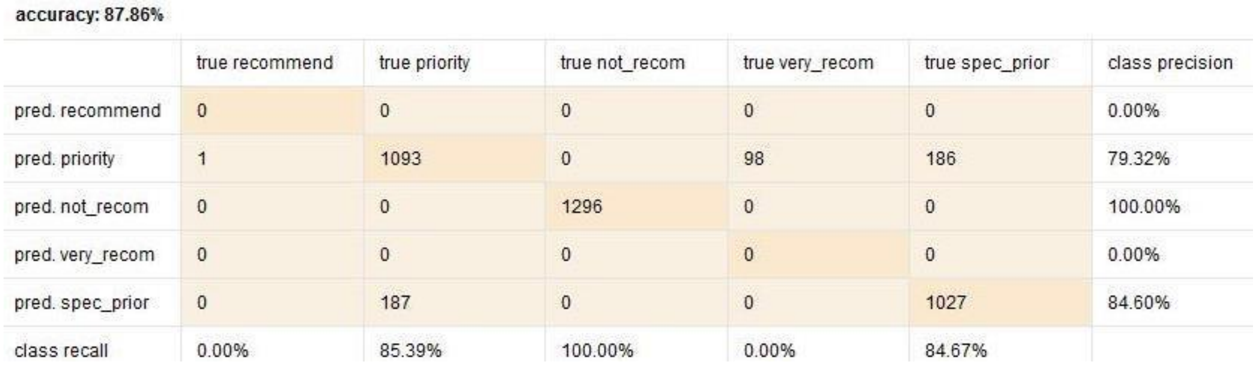
Table 1. First row of the labeled example set

Rank	Prediction(rank)	Confidence (recommend)	confidence (priority)	confidence (not_recom)	confidence (very_recom)	confidence (spec_prior)	has_nurs	parent	health
recommend	priority	0.003296119	0.78508407	2.43E-08	0.168735691	0.042884092	proper	usual	recommended

Figure 4 First row of the labeled example set

Confidence (recommend) = P (Parents = usual | Rank=recommend) * P (has_nurs=proper | Rank=recommend) * P (health = recommended |Rank=recommend) * Posterior (recommend) =0.003

Table1 shows the first row of the labeled dataset. There is a confidence attribute for each possible value of the label. The label value with the highest confidence is assigned as a predicted value for the example. In this example, the highest confidence (i.e., 0.785) is for label value= priority. Therefore this example is predicted as a priority. Similarly done for all other tuples.



accuracy: 87.86%

	true recommend	true priority	true not_recom	true very_recom	true spec_prior	class precision
pred. recommend	0	0	0	0	0	0.00%
pred. priority	1	1093	0	98	186	79.32%
pred. not_recom	0	0	1296	0	0	100.00%
pred. very_recom	0	0	0	0	0	0.00%
pred. spec_prior	0	187	0	0	1027	84.60%
class recall	0.00%	85.39%	100.00%	0.00%	84.67%	

Figure 5 Accuracy using naive Bayes model

A confusion matrix is shown in Figure 5. It shows that the predictions are highly consistent with the dataset (accuracy =87.86%). The result shows that the simple model can generate a good fit for the nursery dataset.

IV.CONCLUSION

This paper observes the performance of the naive Bayes classification method. Bayesian classification delivers a probabilistic outline for a classification problem. It is one of the modest and popular methods that is used for classification. This method is used for classifying the nursery dataset where the dataset has a large set of features and features values to figure. Naïve Bayes is applied on the nursery dataset and the accuracy od model came 87.86% which is good. One of the main restraint of the model is the hypothesis of self-determining features and zero probability problem. In the future, we will use advanced methods for classification and will perform performance comparison with naïve Bayes.

V. REFERENCES

[1]. McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Representation" (PDF). Retrieved 22 October 2019.

[2]. Piryonesi SM, El-Diraby TE. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. Journal of Transportation Engineering, Part B: Pavements. 2020 Jun 1; 146(2):04020022.

[3]. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009 Aug 26.

[4]. Maron ME. Automatic indexing: an experimental inquiry. Journal of the ACM (JACM). 1961 Jul 1; 8(3):404-17.

[5]. Lewis DD. Representation and learning in information retrieval (Doctoral dissertation, University of Massachusetts at Amherst).

[6]. McCallumzy AK, Nigamy K. Employing EM and pool-based active learning for text classification. InProc. International Conference on Machine Learning (ICML) 1998 Jul (pp. 359-367). Cite seer.

[7]. Bhalla R, Bagga A. A Comparative Analysis of Application of Proposed and the Existing Methodologies on a Mobile Phone Survey. In International Conference on Futuristic Trends in Networks and Computing Technologies 2019 Nov 22 (pp. 107-115). Springer, Singapore.

[8]. Bhalla R, Bagga A. Opinion mining framework using proposed rb-bayes model for text classification. International Journal of Electrical & Computer Engineering (2088-8708). 2019 Feb 1;9(1).

[9]. Bhalla R, Bagga A. A Comparative Analysis of Factor Effecting the Buying Judgement of Smart

Phone. International Journal of Electrical & Computer Engineering (2088-8708). 2018 Oct 1;8.

[10]. Hofmann M, Klinkenberg R, editors. Rapid Miner: Data mining use cases and business analytics applications. CRC Press; 2016 Apr 19.

Cite this article as :

Rajni Bhalla, Amandeep, "Observing Performance of Naive Bayes Classifier on Nursery Dataset", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 3, pp. 91-95, May-June 2022. Available at doi : https://doi.org/10.32628/IJSRSET218410
Journal URL : https://ijsrset.com/IJSRSET218410