International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 12 | Dec 2018 www.irjet.net p-ISSN: 2395-0072

Analysis of Chi-Square Independence Test for Naïve Bayes Feature Selection

Pratik Kadam

Graduate Student, Major in Operations Research, Northeastern University, Boston, Massachusetts

---

Abstract - *Naïve Bayes based upon the Bayes theorem and Conditional Probability is one of the most popular classifiers used for classifying data. However, the accuracy of this algorithm highly depends on the features of the datasets used. Present paper thus focuses on using Chi-Square Independence Test for feature selection at various confidence intervals. Tableau was used for data visualization, Minitab as a statistical tool and RStudio was used for developing the Naïve Bayes Model.*

Key Words: **Naïve Bayes Classifier, Chi-Square Independence Test, Feature Selection, Data Science, Student Performance**

1. INTRODUCTION

Naïve Bayes Classifier has been used extensively for spam filtering & other such predictive modelling applications. [1] Naïve Bayes Classifier is a probabilistic classifier based on Bayes' theorem. It assumes independence between the predictors or features of the dataset. This assumption doesn't always necessarily hold true. However, despite the features being dependent, the Naïve Bayes Classifier offers surprisingly accurate predictions. Hence it is termed as 'Naïve' Bayes Classifier.

However, the accuracy of this classifier is highly dependent on the number of features in the dataset. It is quite unclear what is the effect on the accuracy of the model due to factors such as the quantity & quality of features.

The present paper thus attempts at selecting features that are prominent in determining the outcome using the Chi-Square Test of Independence. The features to be selected are determined using the p-values of the features. The accuracy of the classifier would be then judged for different sets of these features based on the mean & standard deviation of a total fraction of correct predictions made.

1.1 CHI-SQUARE TEST OF INDEPENDENCY

To study the dependence of various factors on the failure proportions of the students, Chi-Square test for independence was chosen as the statistical test to be performed assuming the distribution for the sum of squared normal deviates to be a Chi-Square Distribution. This test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables [2]. Following are the prerequisites to be considered before performing the test:

A. The sampling method is a simple random sampling.
B. The variables under study are each categorical.
C. If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

If the above perquisites are satisfied, the following are the steps to be performed to check the dependency of the features.

Step 1: **State the Hypothesis**

H_0: The considered factor and the failure proportion for that factor are independent
H_1: The considered factor and the failure proportion for that factor are not independent

Thus, the alternative hypothesis states that knowing levels of any one factor can help us determine the chances of failure for the student.

Step 2: **Calculating the Test Statistic**

The Chi-Square value is given by the equation below:
$\chi^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$
where,
O[r,c] = Observed frequency counts of pass/fail students for every level of factor.
E[r,c] = (Nr * Nc) / N = Expected frequency counts of a class of students for every level of factor.
Nr = the Total number of pass/fail students for a factor.
Nc = Total number of students one for a level of the factor.
N = Total sample size.

Step 3: **Defining the decision rule**.

Deciding the significance level for p-values.

Step 4: **Making a Decision.**

If the p-value is less than the significance level, we reject H0 and accept H1. i.e. The variables are dependent on each other.

2. DATASET DESCRIPTION

The dataset chosen for this paper, included student's performances in two different Portuguese schools for Math class [3]. The data attributes include student's grades for 3 different periods, demographic, school, personal and social related features. The data was collected was using mark reports and questionnaires. Following are the attributes of the dataset:

1. **school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. **sex** - student's sex (binary: 'F' - female or 'M' - male)
3. **age** - student's age (numeric: from 15 to 22)
4. **address** - student's home address type (binary: 'U' - urban or 'R' - rural)
5. **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8. **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 -5th to 9th grade, 3 - secondary education or 4 - higher education)
9. **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. **guardian** - student's guardian (nominal: 'mother', 'father' or 'other')
13. **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. **failures** - number of past class failures (numeric: n if 1<=n<3, else 4)
16. **schoolsup** - extra educational support (binary: yes or no)
17. **famsup** - family educational support (binary: yes or no)
18. **paid** - extra paid classes within the course subject (binary: yes or no)
19. **activities** - extra-curricular activities (binary: yes or no)
20. **nursery** - attended nursery school (binary: yes or no)
21. **higher** - wants to take higher education (binary: yes or no)
22. **internet** - Internet access at home (binary: yes or no)
23. **romantic** - with a romantic relationship (binary: yes or no)
24. **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
26. **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
27. **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. **health** - current health status (numeric: from 1 - very bad to 5 - very good)
30. **absences** - number of school absences (numeric: from 0 to 93)
31. **G1** - first period grade (numeric: from 0 to 20)
32. **G2** - second period grade (numeric: from 0 to 20)
33. **G3** - final grade (numeric: from 0 to 20, output target) [3].

An extra attribute named 'success' was further added to the existing dataset. This attribute gave information regarding if the student has passed the class or not. This attribute is our categorization attribute. We checked the accuracy of the Naïve Bayes Classifier for two different types of classification. The two types of classification are as follows [4]:

A. 2-class Classification
Pass: If G3 > 10, else **Fail**

B. 5-class Classification
A: G3 = 16 to 20;
B: G3 = 14 to 15;
C: G3 = 12 to 13
D: G3 = 10 to 11;
F: G3 = 0 to 9.

3. DATA VISUALIZATION AND CHI SQUARE TEST RESULTS

Data was at first visualized using Tableau Server to look for any anomalies within the data. Bar charts were generated to visualize the proportions of different classes of success for different levels of every variable.

Following are some of the visualizations obtained from Tableau Server:

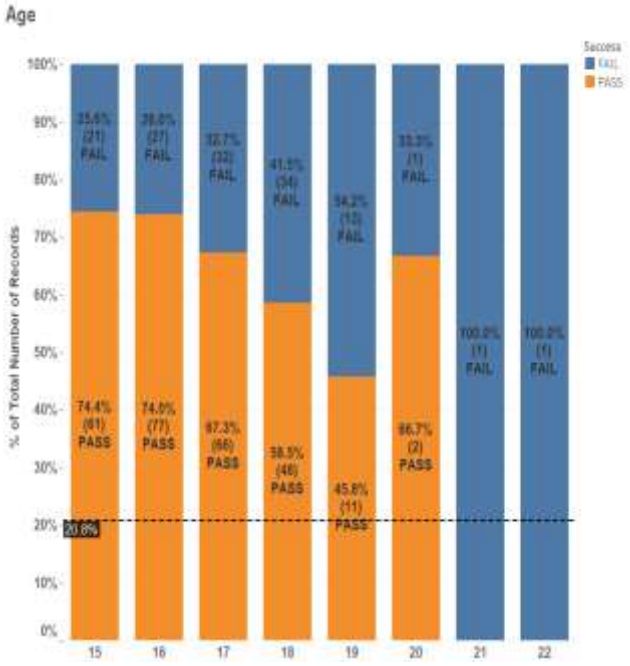

Chart 1: Tableau Server Visualization for % of Total Number of Records for Age and proportion of pass/fail students.

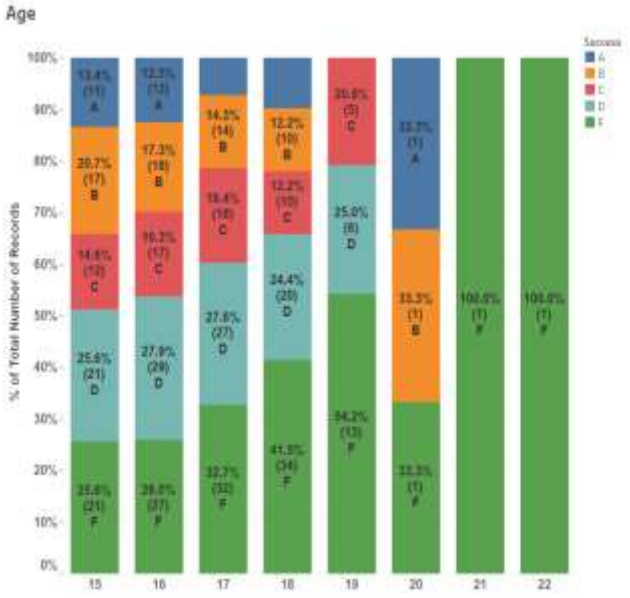

Chart 2: Tableau Server Visualization for % of Total Number of Records for Age and proportion of students with A, B, C, D, F grade.

The data summary obtained was then further used for Chi-Square Independency test. The Chi-Square Independency test was performed using Minitab 18. Following were the results obtained:

Table 1: P-values for Chi-Square Independency test for 2-class Classification.

Features	P-Values
Prev. Class Failures	0
Class Attendance	0
Planning for Higher Education	0.001
Hangouts with Friends	0.002
Age	0.027
School Support	0.038
Relationship Status	0.061
Extra Courses	0.099
Student's Guardian	0.139
Mother's Education	0.143
Family Support	0.146
Sex	0.155
Workday Alcoholic Consumption	0.178
Father's Education	0.193
Reason for Choosing	0.21
Internet Access at Home	0.298
Amount of Free Time	0.334
Mother's Job	0.334
Amount of Study Time	0.387
Parent's Cohabitation Status	0.42
Family Size	0.485
Location	0.521
Quality of Health	0.542
Quality of Family Relationships	0.583
Father's Job	0.674
Travel Time	0.694
Professor	0.714
Attended Nursery	0.74
Weekday Alcoholic Consumption	0.776
Extra Activities	0.807

Table 2: P-values for Chi-Square Independency test for 5-class Classification.

Feature	P-value
Previous Class Failures	0
Class Attendance	0.001
Mother's Education	0.003
School Support	0.005
Planning for Higher Education	0.011
Hangouts with Friends	0.03
Workday Alcoholic Consumption	0.036
Extra Courses	0.049
Relationship Status	0.056
Mother's Job	0.071

Location	0.11
Father's Job	0.158
Sex	0.175
Amount of Study Time	0.202
Father's Education	0.225
Weekday Alcoholic Consumption	0.227
Age	0.237
Family Size	0.296
Internet Access at Home	0.298
Reason for Choosing	0.306
Amount of Free Time	0.339
Student's Guardian	0.354
Travel Time	0.447
Attended Nursery	0.494
Parent's Cohabitation Status	0.569
Quality of Health	0.571
Professor	0.584
Quality of Family Relationships	0.664
Extra Activities	0.693
Family Support	0.722

4. NAÏVE BAYES CLASSIFIER MODELING

Confidence Intervals for the Chi-Squared tests. Confidence Intervals of 75%, 85%, 95% & 99% were decided to be utilized. These datasets were then divided into 80-20% ratio of training and testing datasets respectively using R's random uniform distribution. Modeling was repeated 100 times to mitigate the effect of the random distribution of training and testing datasets.

Furthermore, datasets were constructed considering the effect of the grade attributes on the accuracy. Three separate datasets were built, having no grade attribute, first-period grade, and first two period grades.

The e1071 R package was used for building the Naïve Bayes Model for the dataset. After building the model, it was tested on the test dataset and proportions of correct predictions were obtained. Standard Deviation is the measure of the variance between the repetitions. Following are the results obtained for all types of the datasets:

Table 3: Summary of Correct Predictions for 2-Class Classification.

Data type	Data feature type	Accuracy	Std. Dev.
COMPLETE DATASET	W/O Grades	0.7	0.0435
75%	W/O Grades	0.716	0.0432
85%	W/O Grades	0.714	0.04
95%	W/O Grades	0.734	0.044
99%	W/O Grades	0.724	0.0445
COMPLETE DATASET	G1	0.795	0.0402
75%	G1	0.832	0.0455
85%	G1	0.827	0.0435
95%	G1	0.834	0.0353
99%	G1	0.826	0.04
COMPLETE DATASET	G1 & G2	0.8	0.0422
75%	G1 & G2	0.88	0.0325
85%	G1 & G2	0.884	0.03
95%	G1 & G2	0.885	0.031
99%	G1 & G2	0.885	0.034

Table 4: Summary of Correct Predictions for 5-Class Classification.

Data type	Data feature type	Accuracy	Std. Dev.
COMPLETE DATASET	W/O Grades	0.282	0.0481
75%	W/O Grades	0.305	0.0483
85%	W/O Grades	0.309	0.0483
95%	W/O Grades	0.285	0.0552
99%	W/O Grades	0.272	0.0479
COMPLETE DATASET	G1	0.468	0.0544
75%	G1	0.508	0.0599
85%	G1	0.537	0.0575
95%	G1	0.529	0.0643
99%	G1	0.536	0.0631
COMPLETE DATASET	G1 & G2	0.683	0.0522
75%	G1 & G2	0.692	0.0572
85%	G1 & G2	0.703	0.0573
95%	G1 & G2	0.703	0.0582
99%	G1 & G2	0.695	0.0605

The results had a mean standard deviation of 0.0474 which was acceptable tolerance.

5. CONCLUSIONS

A. For 2-Class Classification

As was expected, highest accuracy was achieved when G1 & G2 both were included in the dataset as G3 was final grade which was heavily dependent on G1 & G2. 88.5% was the maximum accuracy achieved. It was observed that using a Chi-Squared test did help in eliminating unnecessary attributes and increasing the accuracy of the model. However, setting a high confidence interval resulted in over-elimination of variables and decreased the accuracy of the model. Following are the increases in accuracy of the model achieved using the Chi-Squared test for elimination.

Table 5: Summary of Increase in Accuracy for 2-Class Classification.

Data feature type	Data Type (Maximum Accuracy obtained at)	Increase in Accuracy
Without grades	95%	4.86%
G1	95%	4.9%
G1 & G2	95%	10.6%

As we can observe, maximum accuracy was achieved at 95% Confidence Interval with as much as 10.6% increase in accuracy observed.

B. For 5-class Classification

Similarly, for 5-class Classification, highest accuracy was obtained for dataset including G1 & G2. However, unlike 2-class Classification, a maximum accuracy of only 70% was achieved which could possibly be a case of a small dataset and increase in levels of classification. Irrespective of that, Chi-Squared test did improve the accuracy of the model. Maximum accuracy was achieved at 85% Confidence Interval, possibly indicating that model requires a wider dataset for high levels of classification. Following are the increases in the accuracy of the model obtained:

Table 6: Summary of Increase in Accuracy for 5-Class Classification.

Data feature type	Data Type (Maximum Accuracy obtained at)	Increase in Accuracy
Without grades	85%	9.57%
G1	85%	14.74%
G1 & G2	85%	2.93%

We can thus conclude that Chi-Squared is an effective tool in reducing features for building a Naïve Bayes.

REFERENCES

[1] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2] https://stattrek.com/chi-square-test/independence.aspx

[3] https://archive.ics.uci.edu/ml/datasets/student+performance

[4] Paulo Cortez & Alice Silva. "Using Data Mining to Predict Secondary School Student Performance", EUROSIS (2008).

BIOGRAPHY

A Graduate Student at Northeastern University, Boston, Massachusetts with majors in Operations Research.