Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849
e-ISSN: 2548-1398
Vol. 7, No. 4, April 2022

COMPARISON OF INFORMATION GAIN AND CHI-SQUARE SELECTION FEATURES FOR PERFORMANCE IMPROVEMENT OF NAIVE BAYES ALGORITHM ON DETERMINING STUDENTS WITH NO PIP RECIPIENTS AT SMKN 1 BREBES

Magus Sarasnomo, Muljono, M. Arief Soeleman
Master of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
Email: sarasnomo55@gmail.com, muljono@dsn.dinus.ac.id, arief22208@gmail.com

Abstract

All policies of the Smart Indonesia Program (PIP) through the form of the Smart Indonesia Card (KIP) are issued by the government under the auspices of the Ministry of Education and Culture (Kemendikbud) through the National Team for the Acceleration of Poverty Reduction (TNP2K). Helping to alleviate the poor category of students in order to obtain a proper education, prevent children dropping out of school, and fulfill their school needs are the goals of the program. This assistance can be used by students to meet all school needs such as transportation costs to go to school, the cost of buying school supplies, and school pocket money. This study aims to compare the Information Gain and Chi-Square selection features to improve the performance of the Naive Bayes algorithm in determining poor students who are recipients of the Smart Indonesia Program (PIP) at SMKN 1 Brebes, to determine the accuracy of the Naive Bayes, Information Gain and Chi-Square algorithms. and compare the level of accuracy and determine the attributes that affect the accuracy. At this stage, collecting relevant and useful research data, which is collected in the form of literature and data, and processed as research material. Sources of data used in this study in the form of primary data collection and secondary data. The primary data collection technique used in this study was a questionnaire or questionnaire, while the secondary data obtained in this study was through document files. At this stage, preliminary data processing is carried out, the data used is student data of SMKN 1 Brebes in 2021. The initial data collection obtained was 703 data, but not all records were used because they had to go through several stages of initial data processing (data preparation). The results of the Naive Bayes algorithm accuracy of 90.31% with an AUC of 0.967, after the addition of the Information Gain selection feature the accuracy becomes 90.88% with an AUC value of 0.970. The addition of the Information Gain selection feature can help improve the classification performance of the Naive Bayes algorithm even though the accuracy is not maximized. The accuracy of the Naive Bayes algorithm is 90.31% with an AUC of 0.967, after the addition of the Chi-Square selection feature the accuracy becomes 90.88% with an AUC value of 0.970. The accuracy results are not maximized but the addition of the Chi-Square selection feature can also improve the classification performance of the Naive

How to cite: Magus Sarasnomo. et al (2022) Comparison of Information Gain and Chi-Square Selection Features For Performance Improvement of Naive Bayes Algorithm On Determining Students With No PIP Recipients at SMKN 1 Brebes, *Syntax Literate: Jurnal Ilmiah Indonesia, 7*(4).
E-ISSN: 2548-1398
Published by: Ridwan Institute

Bayes algorithm. The accuracy of the Naive Bayes algorithm is 90.31% with an AUC of 0.967, after the addition of the Information Gain selection feature and the Chi-Square selection feature the accuracy becomes 90.88% with an AUC value of 0.970. The results of the same accuracy in the use of the Information Gain and Chi-Square selection features to increase the performance of the Naive Bayes algorithm by 0.57% although the accuracy results are still less than optimal.

Keywords: Information Gain; Chi Square; Algoritma Naïve Bayes; PIP

Introduction

It is hoped that this Smart Indonesia Program (PIP) fund will not happen again for students dropping out of school due to lack of funds. Providing funds for the Smart Indonesia Program (PIP) to underprivileged students from elementary school to high school (Rohaeni & Saryono, 2018). At the Vocational High School (SMK) students will increase every year who come from rich and poor families. Changes in the condition of the community's economic income which sometimes cannot be monitored regularly by the parties concerned will have an impact on the existence of students from wealthy families who are registered as recipients of the Smart Indonesia Card (KIP) and students are not registered as recipients of the Smart Indonesia Card (KIP). from poor families, the policy of the Smart Indonesia Program (PIP) through the Smart Indonesia Card (KIP) has not been fully targeted in equalizing education.

The unavailability of sufficient data and information related to students' family income causes the characteristics of the families of students who are able and cannot be found to be found. Based on this, the authors conducted research on the family income data of students to find the characteristics or characteristics of the family groups of students who were able and unable, as well as proposed new parameters that were more suitable for determining poor students who received PIP at SMKN 1 Brebes, namely orphans or students (Setyawati, 2018). orphans, or orphaned students, where this study aims to increase the level of accuracy.

It is hoped that the research conducted can provide results in the form of useful information in making school decisions, especially those relating to the determination of poor students who receive PIP at SMKN 1 Brebes. Many studies discuss the prediction and determination of student PIP with various data mining algorithm models. In previous studies, prediction techniques and students' PIP determination have been carried out, such as:

Joy Nashar UtamajaYes, Andi Mentari A.P, Siti Masnunah (Utamajaya, Putri, & Masnunah, 2020) in 2020 conducted an analysis of the determination technique to determine prospective PIP scholarship recipients at SDN 023 Penajam by using the Naive Bayes algorithm application model. Some of the criteria used in determining scholarship recipients include: parents' income, parent's occupation, number of dependents, report cards, rank, distance from home to school as well as academic and non-academic achievements. The application of the Naive Bayes method in determining the eligibility of prospective bidikmisi scholarship recipients. The choice of this method is because it is able to study previous case data that used as test data. This research has produced a decision support system application with an accuracy rate of 97.2%.

(RAINI, 2020) in 2020 also examined the evaluation of the implementation of the Smart Indonesia Program at SMA Negeri 1 Sembawa. The use of qualitative descriptive research methods by researchers as a source of primary data and secondary data by collecting data, interviews, documents and library sources. The results of the study can be concluded that the implementation of the Smart Indonesia Program (PIP) at SMAN 1 Sembawa has not run optimally as it should. This can be seen from the number of students who receive PIP funds are students who are not on target, this is partly because the data used in determining candidates PIP receiver is still less accurate.

Several research problems on the characteristics of the family groups of students who are able and unable as a determinant of scholarship recipients or PIP using data mining that have been stated can be carried out an analysis, among others: a. Naive Bayes is a method used to classify a data set (Annur, 2018). b. Feature selection is a technique that is often used in pre-processing data mining by reducing the number of features involved in determining a target class value and reducing irrelevant features (Djatna & Morimoto, 2008).

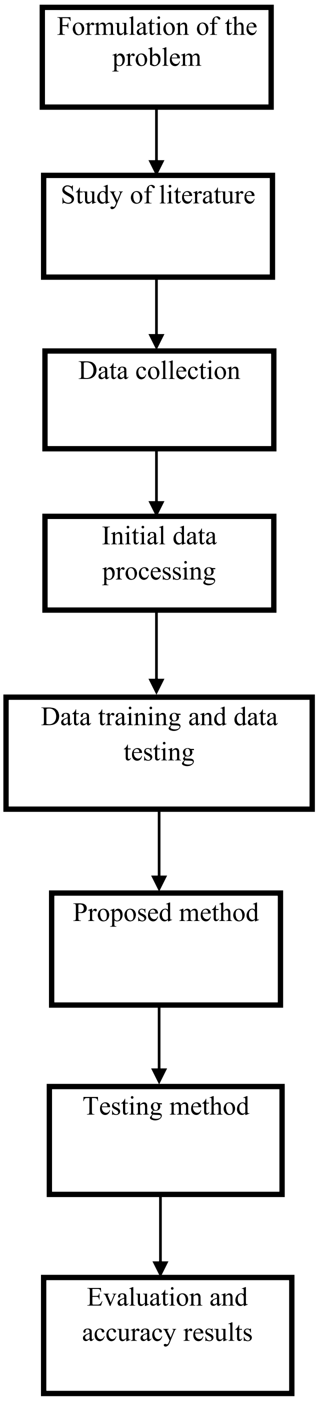
Seeing the ability of data mining with the Naive Bayes algorithm in classifying as well as Information Gain and Chi-Square in feature selection capabilities, the author wants to do comparative research of the two feature selection methods with the Naive Bayes algorithm for determining poor students who are PIP recipients at SMKN 1 Brebes (Betesda, 2020).

In previous studies using Naive Bayes performance in conducting an analysis of determination techniques to determine prospective PIP recipients. In this research on determining poor students who are PIP recipients at SMKN 1 Brebes, we will use an approach with a comparison of Information Gain and Chi-Square selection features to improve the performance of Naive Bayes, while the parameters used are 14 parameters.

The data obtained through the mining process is used to model the algorithm to be used. Result of model This is used to determine the characteristics of the family group of students who are able and unable so that it can determine students who cannot afford PIP recipients at SMKN 1 Brebes (Setiawan, 2017). This study aims to compare the Information Gain and Chi-Square selection features to improve the performance of the Naive Bayes algorithm in determining poor students who are recipients of the Smart Indonesia Program (PIP) at SMKN 1 Brebes and determine the accuracy of the Naive Bayes Algorithm, Information Gain and Chi-Square as well as compare their accuracy. This research can be useful to minimize the ability of students to get PIP, help vocational high schools to pay more attention to the characteristics of the family groups of students who are able and unable so that students who cannot afford to get PIP and help vocational high schools to determine what method should be applied so that students unable to get PIP (Uriyalita & Syahrodi, 2020).

Research Method

Research stages are the steps that will be taken by researchers in providing an overview and ease of conducting a research. Systematically, the steps in this study are presented in Figure 1.


Formulation of the problem
Study of literature
Data collection
Initial data processing
Data training and data testing
Proposed method
Testing method
Evaluation and accuracy results

Sources of data used in this study in the form of primary data collection and secondary data. Primary data in the form of data about the attributes of student data biodata and their families. Secondary data in the form of class division data taken from school archives (Maulidi, 2016).

The primary data collection technique used in this study was a questionnaire or questionnaire, while the secondary data obtained in this study was through document files. The data collected is 703 data with attributes as shown in the following table:

Table 1
Number of Attributes.

Attribute Name	Information
Student's name	Student's Full Name
Gender	Man
	Woman
Class	XI
	XII
Domicile	In the city
	Out of town
Average Report Score	Value < 80
	Value > = 80
Report Ratings	between 1 - 5
	between 6 - 15
	between 16 - 36
Distance to School	Distance < 1 Km
	Distance > = 1 Km
Academic achievement	Yes
	Not
Non-Academic Achievements	Yes
	Not
Parents' job	Laborer
	Fisherman
	Non civil servant
	Farmer
	civil servant
	entrepreneur
Parents' Income	Income < 1,000,000
	Income > 2,000,000
	Income between 1,000,000 - 2,000,000
Number of Family Dependents	Quantity < 2 people
	Quantity > = 2 people
Student Status in Family	Orphans
	Not an orphan
Get PIP	Yes
	Not

Result and Discussion

Experiments that will be carried out by researchers, namely experiments with the Naive Bayes method, experiments with the Naive Bayes and Information Gain methods, experiments with the Naive Bayes and Chi-Square methods, and experiments with the Naive Bayes method and combining the results of the Information Gain and Chi-Square selection features. .

Calculation of Conditional Posterior Probability

Calculating using the Naive Bayes method to determine whether Selvi is in the Yes or No category as a PIP recipient if it is known that the test data or testing from Selvi are:

Name = Selvi
Gender = Female
Class = XI
Domicile = In the city
The average value of the report card = Value < 80
Report card rating = between 16 - 36
Distance to school = Distance > = 1 Km
Academic achievement = No
Non-academic achievement = Yes
Parent's occupation = Labor
Parent's income = Income between 1,000,000 - 2,000,000
Number of dependents = Number > = 2 people
Student status in the family = Not an orphan
Getting PIP = ???

Calculation stage 1:
Based on equation (2):
P(E)=x/n

Calculation stage 2:
The next calculation is continued based on equation (1):
P(x┃y)=(P(y┃x).P(x))/(P(y))

Calculation stage 3:
Multiply all the results of the Yes and No variables.
P(X┃Student=Yes)= 0.797 x 0.437 x 0.975 x 0.294 x 0.417 x 0.888 x 0.506 x 0.118 x 0.437 x 0.166 x 0.863 x 0.818 = 6.716
P(X┃Student=No)= 0.825 x 0.578 x 0.981 x 0.297 x 0.365 x 0.677 x 0.490 x 0.106 x 0.251 x 0.529 x 0.502 x 0.985=6.586

Calculation stage 4:
Result (P|Yes) = 6.716
Result (P|No) = 6.586

Compare the results of the Yes and No classes because the result (P|Yes) is greater than (P|No) then the decision status is Selvi "Yes" to get PIP.

Table 2
Confusion Matrix Calculation in Naive Bayes

Class Actual	Positive Prediction	Negative Prediction	Total
Positive	TP	FN	P
Negative	FP	TN	N
Total	TP + FP	FN + TN	P + N

A model that is trained to predict whether a YES student gets PIP or Does not get PIP, assuming based on table 1 it is known that the total number of students 702 with 439 YES students getting PIP and 263 students Not getting PIP, the Confusion Matrix after calculations using rapidminer studio results in :

- True Positive (TP) which predicts Yes to get PIP and it is true that the student Yes gets PIP as many as 390 students.
- True Negative (TN) which predicts not getting PIP and it is true that the student does not get PIP as many as 244 students.
- False Positive (FP) which predicts Yes to get PIP and it turns out that the prediction is wrong, in fact 19 students don't get PIP.
- False Negative (FN) which predicts Not getting PIP and it turns out that the prediction is wrong, it turns out Yes to get PIP a total of 49 students.

So that after being entered into the table it will look like in table 3.

Table 3
Confusion Matrix results table using Rapidminer Studio

	true No	true Yes
pred. No	244	49
pred. Yes	19	390

Order of information gain value results The information gain value for each attribute is then sorted from the largest value to the lowest value, so that the results are as in table 4 as follows:

Table 4
Table of Results Ranking of Information Gain Value Values

Attribute	Information Gain Value	Rangking
Parents' Income	0,29562	1
Number of Family Dependents	0,10992	2
Student Status in Family	0,05748	3
Parent's Job	0,05177	4
Distance to School	0,04754	5
Class	0,01341	6
Report Ratings	0,00191	7
Gender	0,00085	8
Domicile	0,00028	9
Non-Academic Achievements	0,00024	10
Academic achievement	0,00016	11
Average Report Score	0,00001	12

Confusion Matrix calculation on Naive Bayes and information gain In table 4 the attributes ranked 1 to 6 are taken for the Confusion Matrix calculation, while the other attributes are not used. It is known that the total number of students 702 with 439 YES students getting PIP and 263 students Not getting PIP, After calculating with rapidminer studio it resulted:

- True Positive (TP) which predicts Yes to get PIP and it is true that the student Yes gets PIP as many as 393 students.
- True Negative (TN) which predicts not getting PIP and it is true that the student does not get PIP as many as 245 students.
- False Positive (FP) which predicts Yes to get PIP and it turns out that the prediction is wrong, in fact 19 students don't get PIP.
- False Negative (FN) which predicts Not getting PIP and it turns out that the prediction is wrong, it turns out Yes to get PIP a total of 49 students.

So that after being entered into the table it will look like in table 5.

Table 5
Confusion Matrix results using rapidminer studio

	true No	true Yes
pred. No	245	49
pred. Yes	19	393

The following are contingency tables of the attributes:
Gender Contingency Table
Based on table 1, it is known:
Number of male gender = 135, yes = 89, no = 46
Total female gender = 567, yes = 350, no = 217
Total column Oi (male) = 135, Oi (yes) = 89 and Oi (no) = 46
Total column Oi (female) = 567, Oi (yes) = 350 and Oi (no) = 217
Total row gender Oi (yes) = yes male + yes female = 89+350 = 439
Total row gender Oi (no) = not male + not female = 46+217 = 263
Total gender Oi (yes and no) = 439 + 263 = 702
Using the formula, the value for Ei is calculated for each attribute
Ei= (total row x total column)/(total total)
Ei(Yes,Male)= (439 x 135)/702=84,423
Ei(No,Male)= (263 x 135)/702=50,577
Ei(Yes,Female)= (439 x 567)/702=354,577
Ei(No,Female)= (263 x 567)/702=212.423
Total Ei gender (yes male and yes female) = 84,423 + 354,577 = 439,000
Total Ei gender (not male and not female) = 50,577 + 212,423 = 263,000
Total all gender Ei (yes and no) = 439,000 + 263,000 = 702
Then these values are entered into the sex contingency table as follows:

Table 6
Gender Contingency

Contingency Table						
Get PIP	Gender				Total	
	Male		Female			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	89	84,423	350	354,577	439	439,000
No	46	50,577	217	212,423	263	263,000
TOTAL	135	135	567	567	702	702

Using the formula, the Value for calculated X^2 hitung gender

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(89-84{,}423)^2}{84{,}423} + \frac{(350-354{,}577)^2}{354{,}577} + \frac{(46-50{,}577)^2}{50{,}577} + \frac{(217-212{,}423)^2}{212{,}423} = 0{,}820$

$X^2_{hitung} \leq X^2_{tabel}$, 0,820 ≤ 3,841, then H0 is accepted. No, there is a relationship between gender and getting PIP.

Table 7
Class Contingency

Contingency Table						
Get PIP	Class				Total	
	XI		XII			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	192	215,123	247	223,877	439	439,000
No	152	128,877	111	134,123	263	263,000
TOTAL	344	344	358	358	702	702

Using the formula, the Value for calculated $X^2\ hitung$ class

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(192-215{,}123)^2}{215{,}123} + \frac{(247-223{,}877)^2}{223{,}877} + \frac{(152-128{,}877)^2}{128{,}877} + \frac{(111-134{,}123)^2}{134{,}123} = 13{,}008$

13,008 > 3,841, then H0 is rejected. There is an effect of the relationship between class and Get PIP.

Table 8
Domicile Contingency

Contingency Table						
Get PIP	Domicile				Total	
	In the city		Out of town			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	428	428,994	11	10,006	439	439,000
No	258	257,006	5	5,994	263	263,000
TOTAL	686	686	16	16	702	702

Using the formula, the Value for calculated X^2 *hitung* Domicile

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(428-428{,}994)^2}{428{,}994} + \frac{(11-10{,}006)^2}{10{,}006} + \frac{(258-257{,}006)^2}{257{,}006} + \frac{(5-263)^2}{263} = 0{,}270$

0,270 ≤ 3,841, then H0 is accepted. No, there is a relationship between domicile and Get PIP.

Table 9
Contingency of Average Value of Report Cards

Contingency Table						
Get PIP	Rapor Average Value				Total	
	Value < 80		Value > = 80			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	129	129,449	310	309,551	439	439,000
No	78	77,551	185	185,449	263	263,000
TOTAL	207	207	495	495	702	702

Using the formula, the Value for calculated X^2 *hitung* Rapor Average Value

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(129-129{,}449)^2}{129{,}449} + \frac{(310-309{,}551)^2}{309{,}551} + \frac{(78-77{,}551)^2}{77{,}551} + \frac{(185-185{,}449)^2}{185{,}449} = 0{,}006$

0,006 ≤ 3,841, then H0 is accepted. No, there is a relationship between the average value of the report card and Get PIP.

Table 10
Contingency of Reporting Ratings

Contingency Table						
Get PIP	Rapor Ranking					
	between 1 - 5		between 6 - 15		between 16 -	
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	62	63,786	194	200,739	18	174,474
No	40	38,214	127	120,261	96	104,526
TOTAL	102	102	321	321	27	279
TOTAL						
Oi	Ei					
439	439,000					
263	263,000					
702	702					

Using the formula, the Value for calculated $X^2\ hitung$ Rapor Ranking

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count =

$$\frac{(62-63{,}786)^2}{63{,}786} + \frac{(194-200{,}739)^2}{200{,}739} + \frac{(183-174{,}474)^2}{174{,}474} + \frac{(40-38{,}214)^2}{38{,}214} + \frac{(127-120{,}261)^2}{120{,}261} + \frac{(96-104{,}526)^2}{104{,}526} = 1{,}849$$

$1{,}849 \leq 5{,}991$, then H0 is accepted. No, there is an effect of the relationship between report cards and Get PIP.

Table 11
Distance to School Contingency.

Contingency Table						
Get PIP	Distance to School				Total	
	Distance < 1 Km		Distance > = 1 Km			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	49	83,798	390	355,202	439	439,000
No	85	50,202	178	212,798	263	263,000
TOTAL	134	134	568	568	702	702

Using the formula, the Value for calculated $X^2\ hitung$ Distance to School

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(49-83{,}798)^2}{83{,}798} + \frac{(390-355{,}202)^2}{355{,}202} + \frac{(85-50{,}202)^2}{50{,}202} + \frac{(178-212{,}798)^2}{212{,}798} = 47{,}669$

47,669 > 3,841, then H0 is rejected. There is a relationship between distance to school and Get PIP.

Table 12
Contingency of Academic Achievement.

Contingency Table						
Get PIP	Academic Achievement.				Total	
	Yes		No			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	217	219,500	222	219,500	439	439,000
No	134	131,500	129	131,500	263	263,000
TOTAL	351	351	351	351	702	702

Using the formula, the Value for calculated $X^2\ hitung$ Academic Achievement.

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(217-219{,}500)^2}{219{,}500} + \frac{(222-219{,}500)^2}{219{,}500} + \frac{(134-131{,}500)^2}{131{,}500} + \frac{(129-131{,}500)^2}{131{,}500} = 0{,}152$

0,152 ≤ 3,841, then H0 is accepted. No, there is a relationship between academic achievement and Get PIP.

Table 13
Contingency of Non-Academic Achievements

Contingency Table						
Get PIP	Non-Academic Achievements				Total	
	Yes		No			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	52	50,028	387	388,972	439	439,000
No	28	29,972	235	233,028	263	263,000
TOTAL	80	80	622	622	702	702

Using the formula, the Value for calculated $X^2\ hitung$ Non-Academic Achievements

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Df (Non-Academic Achievements) = (2-1) x (2-1) = 1

Value X^2 tabel = 3,841

Value X^2 count = $\frac{(52-50{,}028)^2}{50{,}028} + \frac{(387-388{,}972)^2}{388{,}972} + \frac{(28-29{,}972)^2}{29{,}972} + \frac{(235-233{,}028)^2}{233{,}028} = 0{,}234$

0,234 ≤ 3,841, then H0 is accepted. No, there is a relationship between non-academic achievement and Get PIP.

Table 14
Parents' Occupational Contingency.

Contingency Table						
Get PIP	Parents' job					
	Workers		Fisherman		Non PNS	
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	192	161,342	7	9,380	9	10,006
No	66	96,658	8	5,620	7	5,994
TOTAL	258	258	15	15	16	16

Contingency Table						
Get PIP	Parents' job					
	Farmer		PNS		entrepreneur	
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	55	55,031	1	11,882	175	191,359
No	33	32,969	18	7,118	131	114,641
TOTAL	88	88	19	19	306	306

TOTAL	
Oi	Ei
439	439,000
263	263,000
702	702

Using the formula, the Value for calculated $X^2\ hitung$ Parents' job

47,766 > 11,070, then H0 is rejected. There is an effect of the relationship between parents' work and Get PIP.

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(192-161{,}342)^2}{161{,}342} + \frac{(7-9{,}380)^2}{9{,}380} + \frac{(9-10{,}006)^2}{10{,}006} + \frac{(66-96{,}658)^2}{96{,}658} + \frac{(8-5{,}620)^2}{5{,}620} + \frac{(7-5{,}994)^2}{5{,}994} + \frac{(55-55{,}031)^2}{55{,}031} + \frac{(1-11{,}882)^2}{11{,}882} + \frac{(175-191{,}359)^2}{191{,}359} + \frac{(33-32{,}969)^2}{32{,}969} + \frac{(18-7{,}118)^2}{7{,}118} + \frac{(131-114{,}641)^2}{114{,}641} =$ 47,766

Table 15
Contingency of Parents' Income

Contingency Table								
Get PIP	Parents' Income						TOTAL	
	Income < 1.000.000		Income > 2.000.000		Income between 1.000.000 - 2.000.000			
	Oi	Ei	Oi	Ei	Oi	Ei	Oi	Ei
Yes	365	267,652	1	38,772	73	132,575	439	439,000
No	63	160,348	61	23,228	139	79,425	263	263,000
TOTAL	428	428	62	62	212	212	702	702

Using the formula, the Value for calculated $X^2\ hitung$ Parents' Income

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count =

$$\frac{(365-267{,}652)^2}{267{,}652} + \frac{(1-38{,}772)^2}{38{,}772} + \frac{(73-132{,}575)^2}{132{,}575} + \frac{(63-160{,}348)^2}{160{,}348} + \frac{(61-23{,}228)^2}{23{,}228} + \frac{(139-79{,}425)^2}{79{,}425} = 264{,}186$$

264,186 > 5,991, then H0 is rejected. There is an effect of the relationship between parents' income and getting PIP.

Table 16
Contingency of Number of Family Dependents

Contingency Table						
Get PIP	Number of Family Dependents				Total	
	Number of Family Dependents < 2 people		Number of Family Dependents > = 2 people			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	60	119,443	379	319,557	439	439,000
No	131	71,557	132	191,443	263	263,000
TOTAL	191	191	511	511	702	702

Using the formula, the Value is calculated for the number of family dependents

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count =

$$\frac{(60-119{,}443)^2}{119{,}443} + \frac{(379-319{,}557)^2}{319{,}557} + \frac{(131-71{,}557)^2}{71{,}557} + \frac{(132-191{,}443)^2}{191{,}443} = 108{,}477$$

108,477 > 3,841, then H0 is rejected. There is a relationship between the number of dependents in the family and getting PIP.

Table 17
Contingency of Student Status in Families

Contingency Table						
Get PIP	Student Status in Families				Total	
	Orphans/Orphans		Not Orphans			
	Oi	Ei	Oi	Ei	Oi	Ei
Yes	80	52,530	359	386,470	439	439,000
No	4	31,470	259	231,530	263	263,000
TOTAL	84	84	618	618	702	702

Using the formula, Value is calculated for the status of students in the family

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Value X^2 count = $\frac{(80-52{,}530)^2}{52{,}530} + \frac{(359-386{,}470)^2}{386{,}470} + \frac{(4-31{,}470)^2}{31{,}470} + \frac{(259-231{,}530)^2}{231{,}530} = 43{,}556$

43,556 > 3.841, then H0 is rejected. There is an effect of the relationship between student status in the family and getting PIP.

Order of results Value Chi Square

The chi-square value for each attribute is then sorted from the largest Yesng Value to the lowest value, so that the results are as shown in table 4.5 as follows:

Table 18
Table of results for the ranking of Value chi-square

Attribute	**Value Chi-Square**	**Rangking**
Parents' Income	264,186	1
Number of Family Dependents	108,477	2
Parents' job	47,766	3
Distance to School	47,669	4
Student Status in Family	43,556	5
Class	13,008	6
Report Ratings	1,849	7
Gender	0,820	8
Domicile	0,270	9
Non-Academic Achievements	0,234	10
Academic achievement	0,152	11
Report Average Value	0,006	12

- Confusion Matrix calculations on Naive Bayes and chi-square
  In table 18 the attributes ranked 1 to 6 are taken to calculate the Confusion Matrix, while the other attributes Yes No are used. It is known that the total number of students 702 with 439 YES students getting PIP and 263 students No getting PIP, After calculating with rapidminer studio it resulted:
- True Positive (TP) Yesitu predicts Yes to get PIP and it is true that the student Yes gets PIP as many as 393 students.
- True Negative (TN) Yesitu predicts No to get PIP and it is true that the student No. gets PIP as many as 245 students.
- False Positive (FP) Yes, it predicts Yes to get PIP and yes, it predicts wrongly, Yesta No gets PIP for 18 students.
- False Negative (FN) Yesitu predicts No to get PIP and yesta predicts wrongly, Yesta turns to get PIP as many as 46 students.
  So that after being entered into the table it will look like in table 19 as follows:

Table 19
Confusion Matrix results using rapidminer studio

	true No	true Yes
pred. No	245	46
pred. Yes	18	393

Based on the confusion matrix table above, the performance of using the Naive Bayes and chi-square classification method can be measured by calculating the value of accuracy, precision and recall.

$$accuracy = \frac{TP+TN}{P+N} = \frac{393+245}{439+263} = \frac{638}{702} = 90{,}88\%$$

$$precision = \frac{TP}{TP+FP} = \frac{393}{393+18} = \frac{394}{411} = 95{,}62\%$$

$$recall = \frac{TP}{TP+FN} = \frac{TP}{P} = \frac{393}{439} = 89{,}52\%$$

Confusion Matrix calculation in Naive Bayes and the results of the information gain and chi-square selection features

The results of combining the Yesng attribute are used in the results of the information gain and chi-square selection features. It is known that the total number of students 702 with 439 YES students getting PIP and 263 students No getting PIP, After calculating with rapidminer studio it resulted:

- True Positive (TP) Yesitu predicts Yes to get PIP and it is true that the student Yes gets PIP as many as 393 students.
- True Negative (TN) Yesitu predicts No to get PIP and it is true that the student No. gets PIP as many as 245 students.
- False Positive (FP) Yes, it predicts Yes to get PIP and yes, it predicts wrongly, Yesta No gets PIP for 18 students.
- False Negative (FN) Yesitu predicts No to get PIP and yes, Yesta predicts wrongly, turns Yesta Yes gets 46 students PIP.

So that after being entered into the table it will look like in table 20 as follows:

Table 20
Confusion Matrix results using rapidminer studio

	true No	true Yes
pred. No	245	46
pred. Yes	18	393

Based on this research can discuss the evaluation of the accuracy of the results of the experimental method to be compared as follows:

Table 21
Attributes of the Naive Bayes Algorithm

Naive Bayes Algorithm
Attribute
Student's name
Getting PIP (Label)
Gender
Class
Domicile
Report Average Value
Report Ratings
Distance to School
Academic achievement
Non-Academic Achievements
Parents' Jobs
Parents' Income
Number of Family Dependents
Student Status in Family

In the Naive Bayes algorithm, all attributes are used for calculations.

Table 22
Information Gain Selection Feature Attributes

Information Gain		
Attribute	**Value**	**Rangking**
Student's name		
Getting PIP (Label)		
Parents' Income	0,29562	1
Number of Family Dependents	0,10992	2
Student Status in Family	0,05748	3
Parent's Job	0,05177	4
Distance to School	0,04754	5
Class	0,01341	6
Report Ratings	0,00191	7
Gender	0,00085	8
Domicile	0,00028	9
Non-Academic Achievements	0,00024	10
Academic achievement	0,00016	11
Report Average Value	0,00001	12

Attributes with low values were not used, such as report card ratings, gender, domicile, non-academic achievement, academic achievement and the average value of report cards because they had no effect on the level of accuracy after being analyzed using rapidminer studio. Only 6 attributes are used, as shown in table 24 below:

Table 24
Attributes of Information Gain Selection Feature Results Used

Information Gain		
Attribute	**Value**	**Rangking**
Student's name		
Getting PIP (Label)		
Parents' Income	0,29562	1
Number of Family Dependents	0,10992	2
Student Status in Family	0,05748	3
Parent's Job	0,05177	4
Distance to School	0,04754	5
Class	0,01341	6

Table 25
Attributes of the Chi Square Selection Feature

Chi-Square		
Attribute	**Value**	**Rangking**
Student's name		
Getting PIP (Label)		
Parents' Income	264,186	1
Number of Family Dependents	108,477	2
Parents' Jobs	47,766	3
Distance to School	47,669	4
Student Status in Family	43,556	5
Class	13,008	6
Report Ratings	1,849	7
Gender	0,820	8
Domicile	0,270	9
Non-Academic Achievements	0,234	10
Academic achievement	0,152	11
Average Report Score	0,006	12

Attributes with low values were not used, such as report card ratings, gender, domicile, non-academic achievement, academic achievement and the average value of report cards because they had no effect on the level of accuracy after being analyzed using rapidminer studio. Only 6 attributes are used, as shown in table 26 below:

Table 26
Attributes of the Chi Square Selection Feature Results used

Chi-Square		
Attribute	**Value**	**Rangking**
Student's name		
Getting PIP (Label)		
Parents' Income	264,186	1
Number of Family Dependents	108,477	2

Parents' job	47,766	3
Distance to School	47,669	4
Student Status in Family	43,556	5
Class	13,008	6

Table 27
Attributes of the Merged Information Gain and Chi-Square Selection Feature Features used

Attributes	
Information Gain	**Chi-Square**
Student's name	Student's name
Getting PIP (Label)	Getting PIP (Label)
Parents' Income	Parents' Income
Number of Family Dependents	Number of Family Dependents
Student Status in Family	Parents' job
Parents' Jobs	Distance to School
Distance to School	Student Status in Family
Class	Class

After the Information Gain and Chi Square attributes are combined, it turns out that the attributes that have an influence on the level of accuracy after being analyzed using rapidminer studio remain the same, namely Student Name, Obtaining PIP (Label), Parental Income, Number of Family Dependents, Student Status in the Family, Occupation of Parents , Distance to School and Class.

Table 28
Results of method comparison

Algorithm	Accuracy	AUC (optimistic)
Naive Bayes	90,31 %	0.967
Naive Bayes and *Information Gain*	90.88%	0.970
Naive Bayes and *Chi-Square*	90.88%	0.970
Naive Bayes, Information Gain and *Chi-Square*	90.88%	0.970

Based on the results of the calculations in this study, the Naive Bayes classification with information gain and chi-square selection features resulted in higher accuracy and AUC values than the Naive Bayes classification without any selection features added. Naive Bayes accuracy value is 90.31% with AUC of 0.967. In this study, the performance of naive bayes increased by 0.57% after using the information gain and chi-square selection features. After doing the calculations in this study, it turns out that the accuracy and AUC results show the same value, namely the accuracy value of 90.88% and the AUC value of 0.970 in the classification of the naive bayes algorithm with information gain and the naive bayes algorithm with chi-square and naive bayes with information gain and chi-square.

Conclussion

This study can be concluded that the accuracy of the Naive Bayes algorithm is 90.31% with an AUC of 0.967, after the addition of the Information Gain selection feature the accuracy becomes 90.88% with an AUC value of 0.970. The addition of the Information Gain selection feature can help improve the classification performance of the Naive Bayes algorithm even though the accuracy is not maximized. The accuracy of the Naive Bayes algorithm is 90.31% with an AUC of 0.967, after the addition of the Chi-Square selection feature the accuracy becomes 90.88% with an AUC value of 0.970. Accuracy results are not maximized but the addition of the Chi-Square selection feature can also improve the classification performance of the Naive Bayes algorithm. The accuracy of the Naive Bayes algorithm is 90.31% with an AUC of 0.967, after the addition of the Information Gain selection feature and the Chi-Square selection feature the accuracy becomes 90.88% with an AUC value of 0.970. The results of the same accuracy in the use of the Information Gain and Chi-Square selection features to increase the performance of the Naive Bayes algorithm by 0.57% although the accuracy results are still less than optimal.

BIBLIOGRAFI

Annur, Haditsah. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes. *Ilkom Jurnal Ilmiah*, *10*(2), 160–165. Google Scholar

Betesda, Betesda. (2020). Peningkatan Optimasi Sentimen Dalam Pelaksanaan Proses Pemilihan Presiden Berdasarkan Opini Publik Dengan Menggunakan Algoritma Naïve Bayes Dan Paricle Swarm Optimization. *Jsi (Jurnal Sistem Informasi) Universitas Suryadarma*, *7*(2), 101–114. Google Scholar

Djatna, Taufik, & Morimoto, Yasuhiko. (2008). Attribute Selection For Numerical Databases That Contain Correlations. *Int. J. Softw. Informatics*, *2*(2), 125–139. Google Scholar

Maulidi, Achmad. (2016). Pengertian Data Primer Dan Data Sekunder. *Online),(Http://Www. Kanalinfo. Web. Id/2016/10/Pengertian-Data-Primer-Dan-Data-Sekunder. Html, Diakses 6 Maret 2017*. Google Scholar

Raini, Agustia Reframamice. (2020). *Pengelolaan Dana Desa Untuk Pemberdayaan Masyarakat*. Google Scholar

Rohaeni, N. Eni, & Saryono, Oyon. (2018). Implementasi Kebijakan Program Indonesia Pintar (Pip) Melalui Kartu Indonesia Pintar (Kip) Dalam Upaya Pemerataan Pendidikan. *Indonesian Journal Of Education Management & Administration Review*, *2*(1), 193–204. Google Scholar

Setiawan, Rony. (2017). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru (Studi Kasus: Politeknik Lp3i Jakarta). *Jurnal Lentera Ict*, *3*(1), 76–92. Google Scholar

Setyawati, Saras. (2018). *Efektivitas Program Kartu Indonesia Pintar (Kip) Bagi Siswa Smk Di Kecamatan Jeruklegi Kabupaten Cilacap (Studi Permendikbud No. 12 Tahun 2015 Tentang Program Indonesia Pintar)*. Iain Purwokerto. Google Scholar

Uriyalita, Fitroh, & Syahrodi, Jamali. (2020). Evaluasi Program Indonesia Pintar (Pip) Telaah Tentang Aksesibilitas, Pencegahan Dan Penanggulangan Anak Putus Sekolah Di Wilayah Urban Fringe Harjamukti, Cirebon. *Edum Journal*, *3*(2), 179–199. Google Scholar

Utamajaya, Joy Nashar, Putri, Andi Mentari Awalia, & Masnunah, Siti. (2020). Penerapan Algoritma Naỉve Bayes Untuk Penentuan Calon Penerima Beasiswa Pip Pada Sdn 023 Penajam. *J-Sim: Jurnal Sistem Informasi*, *3*(1), 11–17. Google Scholar