Classification of Indonesian Smart Card Scholarship Recipients with Principal Component Analysis using the Naive Bayes and Decision Tree Methods Case Study: Stie Pariwisata API Yogyakarta

Bowo Hirwono
Universitas Teknologi Yogyakarta
Jl. Siliwangi (Ringroad Utara)
Sleman, Yogyakarta, Indonesia

Suhirman
Universitas Teknologi Yogyakarta
Jl. Siliwangi (Ringroad Utara)
Sleman, Yogyakarta, Indonesia

ABSTRACT

This research is motivated by the complexity of problems in the selection process for Smart Indonesia Card (KIP) scholarship recipients at STIE Wisata API Yogyakarta. Even though KIP is an important means of increasing access to education for poor families, conventionalmethods of conducting selection cause delays and uncertainty in the validity of the results. In an effort to increase the efficiency and accuracy of selection, this research proposes the application of the Naive Bayes and Decision Tree algorithms.

Metode penelitian melibatkan implementasi Algoritma *Naive Bayes* dan *Decision Tree* untuk mengklasifikasikan kelayakan penerimaan beasiswa KIP, dengan tambahan penerapan Principal Component Analysis (PCA) guna meningkatkan akurasi hasil klasifikasi. Data penerima beasiswa KIP digunakan sebagai input, memungkinkan penelitian ini untuk menguji dan membandingkan performa kedua Algoritma. Penggunaan PCA sebagai dimensi reduksi diharapkan dapat memberikan kontribusi signifikan terhadap hasil akhir.

The research results show that using the Naive Bayes Algorithm with PCA provides the highest accuracy of 85.19%, while Decision Tree with PCA achieves the highest accuracy of 83.33%. The use of PCA is proven to influence significant differences in accuracy in the two algorithms.

Keywords

Kartu indonesia pintar scholarship, naïve bayes, decision tree, principal component analysis.

1. INTRODUCTION

Education is a means to improve a person's standard of living, through education a person can develop well and purposefully, but poverty is one of the problems for students in pursuing education to a higher level. This also causes many high-achieving children to choose to look for work rather than continue their education at university. With this problem, the government is taking part in solving it, including by providing free educational assistance. Smart Indonesia Program (PIP) through the Smart Indonesia Card (KIP) based on PERMENDIKBUD No. 10 of 2020 is education cost assistance provided by the government through the KIP Kuliah program to provide broad access to learning for students and new students from poor or vulnerable families (Gagan Suganda et al., 2022). The KIP carried out by the government is one form of part in perfecting the Poor Student Assistance Program (BSM). So far, the selection stage carried out for prospective students still uses the conventional method, namely by collecting files directly as a specified requirement. This makes the selection process take quite a long time and the results are not necessarily valid because there are more prospective students who register than the available quota. Therefore, to assist selectors in making decisions, it is deemed necessary to classify the eligibility for receiving KIP scholarships using a data mining approach, namely the Naive Bayes algorithm and decision trees.

Education is a means to improve one's standard of living, through an education a person can develop well and purposefully, but poverty is one of the problems of students in receiving education to a higher level. Not a few of these also cause many outstanding children to choose to find a job rather than continue their education to college. Moreover, if the assumption of conditional independence applies, it can certainly give good results, this algorithm belongs to an old algorithm but is still popular in classification, besides the use of the Naive Bayes method helps in estimating the likelihood that someone will receive a scholarship based on historical data and other related information. After using the naïve bayes algorithm, researchers did the next classification as a comparison using the Decision Tree algorithm for the reason that the Decision Tree algorithm has the advantage of being able to explore hidden information in a large data, divide large data sets into smaller sets and the results of analysis in the form of tree diagrams that are easy to understand and also help in determining the stages of logic that It is needed in the decision making process of scholarship acceptance, so with that this method can help STIE Pariwisata Api Yogyakarta which still uses conventional methods in classifying KIP scholarship acceptance which will be used as a benchmark in the process of providing scholarships to its students.

2. LITERATURE REVIEW AND THEORETICAL FOUNDATIONS

2.1 Literature Review

Based on the results of research conducted by Tempola (2021) regarding the selection of smart Indonesia card (KIP) recipients using a dataset of 150 data, testing training data 1 to 149 and test data 149 to 1 using the value of K = 3 obtained the lowest accuracy of 48% and the highest accuracy of 100%. Based on the test results that have been carried out in applying the K-NN incision method K=3, the accuracy is better than K=5 or K=7 if the data is divided equally for the testing process in terms of accepting the Smart Indonesia Card. However, when compared to the average accuracy of applying the Naive Bayes method, the accuracy of Naive Bayes is higher than k-NN which obtained an average accuracy of 85.66%.

2.1.1 Kartu Indonesia Pintar Scholarship (KIP)

Education is a key factor in alleviating poverty in Indonesian society, therefore education is one of the important aspects to measure the success of a country's development. the government takes an important role in handling these problems, including providing Smart Indonesia Card (KIP) scholarships, smart Indonesia Cards (KIP) is an educational assistance program for students who graduated from high school, vocational school and equivalent from underprivileged families in order to continue their education to the university or academy level Not only funding new students who are accepted, KIP Kuliah also funds Bidikmisi on-going recipients who are currently receiving Bidikmisi scholarships who are currently studying (Fadhli, R. 2021).

2.1.2 Data Mining

Data mining is a technology that can process large volumes of data used by companies to turn raw data into information that is useful for making very important business decisions. Basically, Data Mining has 7 functions, namely Description, Classification, Clustering, Association, Sequencing, Forecasting, and Prediction. Data mining has a purpose as an Explanatory, which is to explain some conditions related to a study. Data mining architecture is a concept that shows the flow of data mining processing starting from retrieving information from the data source to be used, data processing, to the relationship between the data mining system and the user or user (Ardilla, 2021).

2.1.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a method for reducing the dimensions of attributes in the dataset, so that the values formed are very different from the original form. The PCA method is used to summarize the structure of a dataset with many dimensions so that it has a smaller number of variables. The use of PCA in addition to reducing dimensions is that it can be used as a method to test whether each variable in the dataset is related or not related at all (Azmi, 2023).

2.1.4 Decision Tree

A decision tree is a tree-like flow chart in which each internal node represents a test of an attribute, each branch represents the output of the test and leaf node represents classes or class distributions. The topmost node is referred to as the root node. A root node will have multiple edge out but no edge in, an internal node will have one edge in and multiple edge out, whereas a leaf node will only have one edge in without having an edge out. Decision Tree is used to classify an unknown sample of data into existing classes. Algorithm - the algorithm in the Decision Tree. There are many algorithms on this Decision Tree classification. An algorithm is usually developed to improve the performance of an existing algorithm (Qadrini L et al., 2021).

2.1.5 Naïve Bayes

The Bayesian model (BM) is a simple probabilistic model built from Bayes' theorem (or Bayes' rule), which has three main components: before, conditional, and posterior probability. The original concept of Bayes' rule was that the outcome of an event (A) could be predicted based on some observable evidence (B) (Noor et al., 2018). The Naïve Bayes algorithm is also one of the solving procedures contained in elaboration techniques that use simple probability methods based on Bayes' theorem with high independent estimates. In addition, this method also shows high accuracy and speed when used in large databases (Annur, 2018). Another definition says Naive Bayes is a classification with probability and statistical methods developed by British scientist Thomas Bayes, namely predicting future opportunities based on previous experience.

3. RESEARCH METHODOLOGY

3.1 Research Materials

This research took data from students of the 2021 and 2022 batches of STIE Api Tourism Yogyakarta. The amount of data used was 180 students, consisting of 136 students graduating and 44 students not graduating. The composition of the data can be seen in the table below:

Table 3.1 Composition of Datasets Used

Student Status	Sum
Graduating students	136
Students who did not graduate	44
Total	**180**

The data is divided into 2 (two), namely training and testing data. The dataset attributes used in this study consisted of 11 attributes including KIP (smart Indonesia card), KKS (prosperous family card), PKH (family hope program), SKTM (certificate of disability), Diploma, Pass Letter, Electricity Bill, Number of Houses, FC Report Card, Parents' Income, Description (Pass No). The attributes of Electricity Bill, Number of Houses, FC Report Card, while for the Remarks attribute that should contain Pass and not pass is converted to 1 for pass, and 0 for not pass.

Tabel 3.2 Variabel dan Tipe Data

No	Attribute	Information
1	KIP (Smart Indonesia Card)	0 = None 1 = Exist
2	KKS (Prosperous Family Card)	0 = None 1 = Exist
3	PKH (Family Hope Program)	0 = None 1 = Exist
4	SKTM (Certificate of Incapacity)	0 = None 1 = Exist
5	diploma	0 = None 1 = Exist
6	Pass Letter	0 = None 1 = Exist
7	Electricity Bill	0 = None 1 = Exist
8	Total Houses	0 = None 1 = Exist
9	Copy of Report Card	0 = None 1 = Exist
10	Parents' Income	0 = None 1 = Exist
11	Information (pass/ not)	0 = None 1 = Exist

3.2 Research Flow

The following is an explanation of the stages of research in the flow of the classification process that will be carried out from beginning to end, as shown in the picture below:

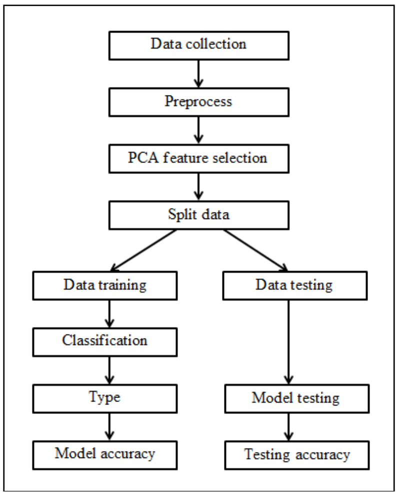


Figure 3.1 Research Process

1. Data Collection
   Dataset sampling by pulling data from the STIE Pariwisata Api Yogyakarta puskom.
2. Preprocess Data.
   After the data is obtained, the next stage is the data preprocessing stage, where at this stage the data will be normalized and divided into 2 (two) parts, namely train data or training data and also testing data or called test data which is used as a model after the training process is complete.
3. Feature Selection.
   Feature selection is a pre-process stage that is useful especially in eliminating all features or information that are considered irrelevant and less effective.
4. Split Data.
   Split Data is done by dividing the data into 2 parts, namely training data and test data, the training data is added to the model to update parameters in the training phase.

Tabel 3.1 Percentage of Train Data and Test Data

No	Percentage of Train Data	Test Data Percentage
1.	50%	50%
2.	55%	45%
3.	60%	40%
4.	65%	35%
5.	70%	30%
6.	75%	25%
7.	80%	20%
8.	85%	15%

5. Testing

At this stage researchers will test a set of existing data using RapidMiner Software, the software testing process is carried out to ensure that the software developed is running properly to produce accuracy values.

4. RESULTS OF RESEARCH AND DISCUSSION

4.1 RapidMiner Implementation on Naïve Bayes Classification Without PCA

The classification of Naïve Bayes in RapidMiner requires multiple panels to be connected. First, the Retrieve data panel to pull the data that has been imported, then connected to the Nominal to Numerical panel which functions to convert category data into numeric values by labeling each category with numbers. The next panel is connected to the Split Data panel which serves to divide training data and testing data. The Split data panel will be linked to two other panels, the Naïve Bayes panel and the Apply Model panel. The Naïve Bayes panel is used to perform training, then the Apply Model panel is used to deploy the formed model. After that, the Apply model panel is connected to the performance panel to display the accuracy results of the experiments conducted. The application of the Naïve Bayes classification to RapidMiner can be shown in Figure 4.1.

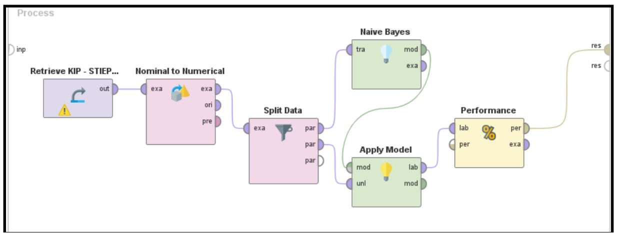
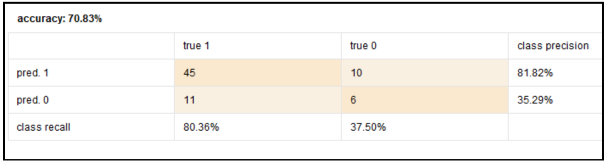


Figure 4.1 Naïve Bayes schematic

The Naïve Bayes classification experiment without PCA was conducted with eight data splits as per Table 3.2 on RapidMiner, the highest accuracy performance results will be shown as in Figure 4.2.


accuracy: 70.83%

	true 1	true 0	class precision
pred. 1	45	10	81.82%
pred. 0	11	6	35.29%
class recall	80.36%	37.50%	

Figure 4.2 Highest Accuracy Test Results - Naive Bayes

The results of the Naïve Bayes classification experiment without PCA are presented in Table 4.1.

Table 4.1 Test Results of Naive Bayes Algorithm Without PCA

No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	33.33%	66.67%	28.57%
2	80%	20%	41.67%	73.33%	39.29%
3	75%	25%	44.44%	72.73%	45.71%
4	70%	30%	55.56%	76.47%	61.90%
5	65%	35%	55.56%	80.00%	57.14%
6	60%	40%	70.83%	81.82%	80.36%
7	55%	45%	67.90%	82.46%	74.60%
8	50%	50%	60.00%	75.76%	71.43%

Table 4.1 shows the results of tests performed with the Naïve Bayes Algorithm without the use of PCA. The experimental results of eight split data found that the lowest accuracy was found in split data 85%:15% with accuracy 33.33%, precision 66.67%, and recall 28.57%, while the highest accuracy results were found in split data 60%:40%, namely accuracy 70.83%, precision 81.82%, and recall 80.36%.

4.2 RapidMiner Implementation on Naïve Bayes Classification with PCA

Naïve Bayes classification in RapidMiner with PCA requires multiple panels connected. First, the Retrieve data panel to pull the data that has been imported, then connected to the Nominal to Numerical panel which functions to convert category data into numeric values by giving labels in the form of numbers in each category. The Nominal to Numerical panel is connected to the PCA panel, and then connected to the Split Data panel which functions to divide training and testing data. The Split data panel will be linked to two other panels, the Naïve Bayes panel and the Apply Model panel. The Naïve Bayes panel is used to perform training, then the Apply Model panel is used to deploy the formed model. After that, the Apply model panel is connected to the performance panel to display the accuracy results of the experiments conducted. The application of the Naïve Bayes classification to RapidMiner can be shown in Figure 4.3.

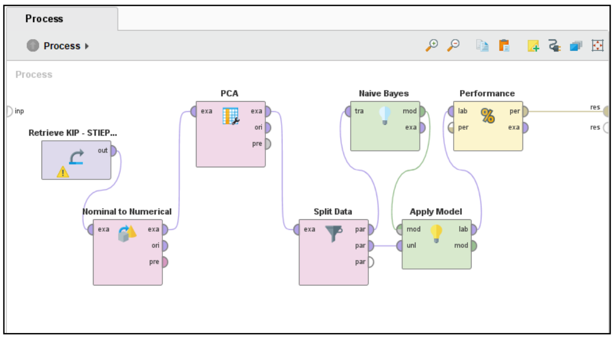
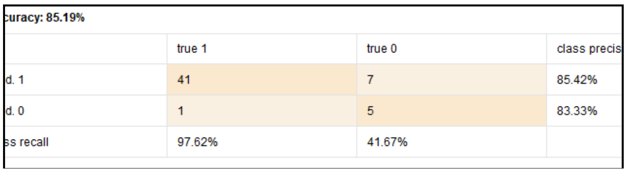


Figure 4.3 Naïve Bayes schematic with PCA

The highest accuracy performance results will be shown as Figure 4.4



curacy: 85.19%

	true 1	true 0	class precis
d. 1	41	7	85.42%
d. 0	1	5	83.33%
ss recall	97.62%	41.67%	

Figure 4.4 Highest Accuracy Test Results - Naive Bayes with

4.2.1 Naïve Bayes Algorithm Test Results with PCA 2 Component (C=2)

The results of the Naïve Bayes classification experiment with the number of PCA components that are 2 (C = 2) are presented in Table 4.2.

Table 4.2 Naive Bayes Algorithm Test Results with PCA (C=2)

Naïve Bayes - PCA 2 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	77.78%	77.78%	100%
2	80%	20%	77.78%	77.78%	100%
3	75%	25%	77.78%	77.78%	100%
4	70%	30%	77.78%	77.78%	100%
5	65%	35%	77.78%	77.78%	100%
6	60%	40%	77.78%	77.78%	100%
7	55%	45%	77.78%	77.78%	100%
8	50%	50%	77.78%	77.78%	100%

Table 4.2 shows the results of tests performed with the Naïve Bayes Algorithm by applying PCA with the number of components C=2. The experimental results of eight split data found that the accuracy of all types of split data was the same, namely 77.78% accuracy, 77.78% precision, and 100% recall

4.2.2 Naïve Bayes Algorithm Test Results with PCA 3 Components (C=3)

The results of the Naïve Bayes classification experiment with the number of PCA components that are 3 (C = 3) are presented in Table 4.3.

Table 4.3 Naive Bayes Algorithm Test Results with PCA (C=3)

Naïve Bayes - PCA 3 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	77.78%	77.78%	100%
2	80%	20%	77.78%	77.78%	100%
3	75%	25%	77.78%	77.78%	100%
4	70%	30%	77.78%	77.78%	100%
5	65%	35%	77.78%	77.78%	100%
6	60%	40%	77.78%	77.78%	100%
7	55%	45%	77.78%	77.78%	100%
8	50%	50%	77.78%	77.78%	100%

Table 4.3 shows the results of tests conducted with the Naïve Bayes Algorithm by applying PCA with the number of components C=3, where the test results with the number of components C=3 are equal to the results of the C=2 test shown in Table 4.2. The experimental results of eight split data found that the accuracy of all types of split data was the same, namely 77.78% accuracy, 77.78% precision, and 100% recall.

4.2.3 Naïve Bayes Algorithm Test Results with PCA 4 Components (C=4)

The results of the Naïve Bayes classification experiment with the number of PCA components which is 4 (C = 4) are presented in Table 4.4.

Table 4.4 Naive Bayes Algorithm Test Results with PCA (C=4)

Naïve Bayes - PCA 4 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	77.78%	77.78%	100%
2	80%	20%	77.78%	77.78%	100%
3	75%	25%	77.78%	77.78%	100%
4	70%	30%	77.78%	77.78%	100%
5	65%	35%	77.78%	77.78%	100%
6	60%	40%	77.78%	77.78%	100%
7	55%	45%	71.60%	77.78%	88.89%
8	50%	50%	77.78%	77.78%	100%

Table 4.4 shows the results of tests performed with the Naïve Bayes Algorithm by applying PCA with the number of components C=4. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 71.60%, precision 77.78& and recall 88.89%, while the highest accuracy was accuracy of 77.78%, precision of 77.78%, and recall of 100% obtained from 7 split data schemes other than split data 55%:45%.

4.2.4 .Naïve Bayes Algorithm Test Results with PCA 5 Components (C=5)

The results of the Naïve Bayes classification experiment with the number of PCA components which is 5 (C = 5) are presented in Table 4.5.

Table 4.5 Test Results of Naive Bayes Algorithm with PCA (C=5)

Naïve Bayes - PCA 5 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	77.78%	77.78%	100%
2	80%	20%	77.78%	77.78%	100%
3	75%	25%	77.78%	77.78%	100%
4	70%	30%	77.78%	77.78%	100%
5	65%	35%	77.78%	77.78%	100%
6	60%	40%	73.61%	77.61%	92.86%
7	55%	45%	71.60%	77.78%	88.89%
8	50%	50%	73.33%	78.05%	91.43%

Table 4.5 shows the results of tests performed with the Naïve Bayes Algorithm by applying PCA with the number of components C=5. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 71.60%, precision 77.78& and recall 88.89%, while the highest accuracy was accuracy of 77.78%, precision 77.78%, and recall 100% obtained from 5 split data schemes, namely in split data 85%:15%, 80%:20%, 75%:25%, 70%:30%, and 60%:40%.

4.2.5 Naïve Bayes Algorithm Test Results with PCA 6 Components (C=6)

The results of the Naïve Bayes classification experiment with the number of PCA components of 6 (C=6) are presented in Table 4.6.

Tabel 4.6 Hasil Uji Algoritma Naive Bayes dengan PCA (C=6)

Naïve Bayes - PCA 6 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	81.48%	80.77%	100%
2	80%	20%	80.56%	81.82%	96.43%
3	75%	25%	80.00%	80.95%	97.14%
4	70%	30%	85.19%	85.42%	97.62%
5	65%	35%	79.37%	83.33%	91.84%
6	60%	40%	81.94%	87.72%	89.29%
7	55%	45%	76.54%	86.67%	82.54%
8	50%	50%	81.11%	84.42%	92.86%

Table 4.6 shows the results of tests performed with the Naïve Bayes Algorithm by applying PCA with the number of components C=6. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 76.54%, precision 86.67& and recall 82.54%, while the highest accuracy was found in split data 70%:30% with an accuracy value of 85.19%, precision 85.42%, and recall 97.62%.

4.2.6 Naïve Bayes Algorithm Test Results with PCA 7 Components (C=7)

The results of the Naïve Bayes classification experiment with the number of PCA components of 7 (C=7) are presented in Table 4.7.

Table 4.7 Naive Bayes Algorithm Test Results with PCA (C=7)

Naïve Bayes - PCA 7 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	81.48%	80.77%	100%
2	80%	20%	80.56%	81.82%	96.43%
3	75%	25%	80.00%	80.95%	97.14%
4	70%	30%	83.33%	83.67%	97.62%
5	65%	35%	80.95%	86.27%	89.80%
6	60%	40%	81.94%	87.72%	89.29%
7	55%	45%	76.54%	89.29%	79.37%
8	50%	50%	81.11%	84.42%	92.86%

Table 4.7 shows the results of tests performed with the Naïve Bayes Algorithm by applying PCA with the number of components C=7. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 76.54%, precision 89.29& and recall 79.37%, while the highest accuracy was found in split data 70%:30% with an accuracy value of 83.33%, precision 83.67%, and recall 97.62%.

4.2.7 Naïve Bayes Algorithm Test Results with PCA 8 Components (C=8)

The results of the Naïve Bayes classification experiment with the number of PCA components which is 8 (C = 8) are presented in Table 4.8.

Table 4.8 Naive Bayes Algorithm Test Results with PCA (C=8)

Naïve Bayes - PCA 8 Component					
No	**Training data**	**Test Data**	**Accuracy**	**Precision**	**Recall**
1	85%	15%	81.48%	80.77%	100%
2	80%	20%	80.56%	81.82%	96.43%
3	75%	25%	80.00%	80.95%	97.14%
4	70%	30%	83.33%	83.67%	97.62%
5	65%	35%	80.95%	84.91%	91.84%
6	60%	40%	79.17%	84.75%	89.29%
7	55%	45%	74.07%	83.87%	82.54%
8	50%	50%	78.89%	83.12%	91.43%

Table 4.8 shows the results of tests performed with the Naïve Bayes Algorithm by applying PCA with the number of components C=8. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 74.07%, precision 83.87& and recall 82.54%, while the highest accuracy was found in split data 70%:30% with an accuracy value of 83.33%, precision 83.67%, and recall 97.62%.

4.3 RapidMiner Implementation on Decision Tree Classification Without PCA

Decision Tree classification in RapidMiner requires multiple panels to be connected. First, the Retrieve data panel to pull the data that has been imported, then connected to the Nominal to Numerical panel which functions to convert category data into numeric values by giving labels in the form of numbers in each category. The next panel is connected to the Split Data panel which serves to divide training data and testing data. The Split data panel will be connected to two other panels, namely the Decision Tree panel and the Apply Model panel. The Decision Tree panel is used to conduct training, and the Apply Model panel is used to deploy the model that has been formed. After that, the Apply model panel is connected to the performance panel to display the accuracy results of the experiments conducted. The application of Decision Tree classification to RapidMiner can be shown in Figure 4.6.

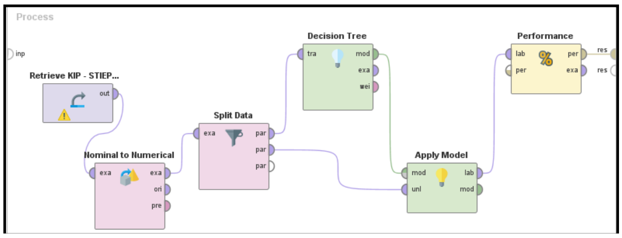
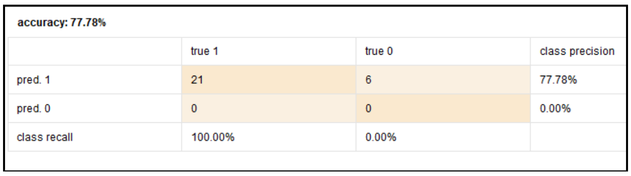


Figure 4.5 Naïve Bayes schematic

The Decision Tree classification experiment without PCA was conducted with eight data splits according to Table 3.2 on RapidMiner, the highest accuracy performance results will be shown as shown in Figure 4.7.


accuracy: 77.78%

	true 1	true 0	class precision
pred. 1	21	6	77.78%
pred. 0	0	0	0.00%
class recall	100.00%	0.00%	

Figure 4.6 Highest Accuracy Test Results - Decision Tree

The results of the Decision Tree classification experiment without PCA are presented in Table 4.9.

Table 4.9 Decision Tree Algorithm Test Results Without PCA

No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	77.78%	77.78%	100.00%
2	80%	20%	77.78%	77.78%	100.00%
3	75%	25%	77.78%	79.07%	97.14%
4	70%	30%	77.78%	77.78%	100.00%
5	65%	35%	77.78%	77.78%	100.00%
6	60%	40%	77.78%	77.78%	100.00%
7	55%	45%	67.90%	76.81%	84.13%
8	50%	50%	76.67%	78.16%	97.14%

Table 4.9 shows the results of tests performed with the Decision Tree Algorithm without the use of PCA. The experimental results of eight split data found that the lowest accuracy was found in split data 55%:45% with accuracy of 67.90%, precision 76.81%, and recall 84.13%, while the highest accuracy results were found in split data 85%:15%, 80%:20%, 70%30%, 65%:35%, and 60%:40% with accuracy values of 77.78%, precision 77.78%, and recall 100%.

4.4 RapidMiner Implementation on Decision Tree Classification with PCA

Decision Tree classification in RapidMiner with PCA requires multiple panels to be connected. First, the Retrieve data panel to pull the data that has been imported, then connected to the Nominal to Numerical panel which functions to convert category data into numeric values by giving labels in the form of numbers in each category. The Nominal to Numerical panel is connected to the PCA panel, and then connected to the Split Data panel which functions to divide training and testing data. The Split data panel will be connected to two other panels, namely the Decision Tree panel and the Apply Model panel. The Decision Tree panel is used to conduct training, and the Apply Model panel is used to deploy the model that has been formed. After that, the Apply model panel is connected to the performance panel to display the accuracy results of the experiments conducted. The application of Decision Tree classification to RapidMiner can be shown in Figure 4.8.

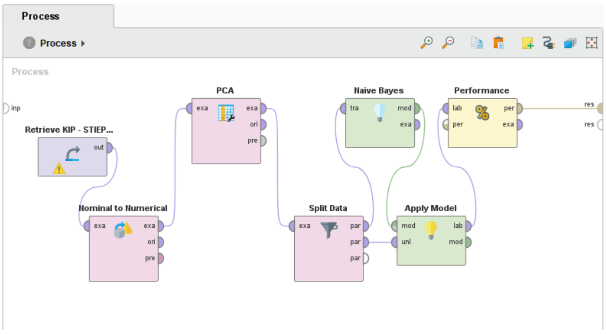
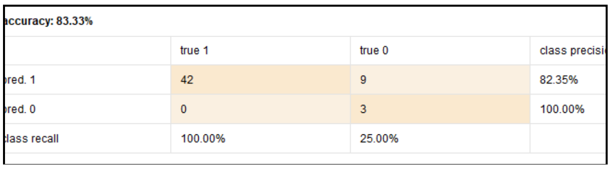


Figure 4.8 Decision Tree schematic with PCA

The highest accuracy performance results will be shown as Figure 4.9.


ccuracy: 83.33%

	true 1	true 0	class precisi
red. 1	42	9	82.35%
red. 0	0	3	100.00%
lass recall	100.00%	25.00%	

Figure 4.7 Highest Accuracy Test Results - Decision Tree with PCA

4.4.1 Decision Tree Algorithm Test Results with PCA 2 Components (C=2)

The results of the Decision Tree classification experiment with the number of PCA components that are 2 (C = 2) are presented in Table 4.10.

Table 4.10 Decision Tree Algorithm Test Results with PCA (C=2)

Decision Tree - PCA 2 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	77.78%	77.78%	100%
2	80%	20%	77.78%	77.78%	100%
3	75%	25%	77.78%	77.78%	100%
4	70%	30%	77.78%	77.78%	100%
5	65%	35%	77.78%	77.78%	100%

Decision Tree - PCA 2 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
6	60%	40%	77.78%	77.78%	100%
7	55%	45%	77.78%	77.78%	100%
8	50%	50%	77.78%	77.78%	100%

Table 4.10 shows the results of tests performed with the Decision Tree Algorithm by applying PCA with the number of components C=2. The experimental results of eight split data found that the accuracy in all types of split data was the same, namely 77.78% accuracy, 77.78% precision, and 100% recall.

4.4.2 Decision Tree Algorithm Test Results with PCA 3 Components (C=3)

The results of the Decision Tree classification experiment with the number of PCA components that are 3 (C = 3) are presented in Table 4.11.

Table 4.11 Decision Tree Algorithm Test Results with PCA (C=3)

Decision Tree - PCA 3 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	77.78%	77.78%	100.00%
2	80%	20%	77.78%	77.78%	100.00%
3	75%	25%	77.78%	77.78%	100.00%
4	70%	30%	77.78%	77.78%	100.00%
5	65%	35%	77.78%	77.78%	100.00%
6	60%	40%	77.78%	77.78%	100.00%
7	55%	45%	74.07%	80.00%	88.89%
8	50%	50%	73.33%	78.05%	91.43%

Table 4.11 shows the results of tests conducted with the Decision Tree Algorithm by applying PCA with the number of components C=3, where the test results with the number of components C=3 found that the lowest accuracy was found in split data 50%:50%, namely accuracy 73.33%, precision 78.05%, and recall 91.43%. The highest accuracy is with a value of 77.78%, precision 77.78%, and 100% recall in split data 85%:15%, 80%:20%, 75%:25%, 70%:30%, 65%:35% and 60%:40%.

4.4.3 Decision Tree Algorithm Test Results with PCA 4 Components (C=4)

The results of the Decision Tree classification experiment with the number of PCA components that are 4 (C = 4) are presented in Table 4.12.

Tabel 4.11 Hasil Uji Algoritma Decision Tree dengan PCA (C=4)

Decision Tree - PCA 4 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	77.78%	77.78%	100.00%
2	80%	20%	77.78%	77.78%	100.00%
3	75%	25%	77.78%	77.78%	100.00%
4	70%	30%	77.78%	77.78%	100.00%
5	65%	35%	77.78%	77.78%	100.00%
6	60%	40%	75.00%	77.94%	94.64%
7	55%	45%	71.60%	77.78%	88.89%
8	50%	50%	73.33%	78.05%	91.43%

Table 4.12 shows the results of tests performed with the Decision Tree Algorithm by applying PCA with the number of components C=4. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 71.60%, precision 77.78& and recall 88.89%, while the highest accuracy was accuracy of 77.78%, precision of 77.78%, and recall of 100% obtained from 6 split data schemes, namely 85%:15%, 80%:20%, 75%:25%, 70%:30%, and 65%:35%.

4.4.4 Decision Tree Algorithm Test Results with PCA 5 Components (C=5)

The results of the Decision Tree classification experiment with the number of PCA components which is 5 (C = 5) are presented in Table 4.13.

Table 4.13 Decision Tree Algorithm Test Results with PCA (C=5)

Decision Tree - PCA 5 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	77.78%	77.78%	100.00%
2	80%	20%	77.78%	77.78%	100.00%
3	75%	25%	77.78%	77.78%	100.00%
4	70%	30%	77.78%	77.78%	100.00%
5	65%	35%	77.78%	77.78%	100.00%
6	60%	40%	75.00%	77.94%	94.64%
7	55%	45%	71.60%	77.78%	88.89%
8	50%	50%	74.44%	79.01%	91.43%

Table 4.13 shows the results of tests performed with the Decision Tree Algorithm by applying PCA with the number of components C=5. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 71.60%, precision 77.78& and recall 88.89%, while the highest accuracy was accuracy of 77.78%, precision 77.78%, and recall 100% obtained from 5 split data schemes, namely in split data 85%:15%, 80%:20%, 75%:25%, 70%:30%, and 65%:35%.

4.4.5 Decision Tree Algorithm Test Results with PCA 6 Components (C=6)

The results of the Decision Tree classification experiment with the number of PCA components of 6 (C = 6) are presented in Table 4.14.

Table 4.14 Decision Tree Algorithm Test Results with PCA (C=6)

Decision Tree - PCA 6 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	81.48%	80.77%	100%
2	80%	20%	80.56%	80.00%	100%
3	75%	25%	80.00%	79.55%	100%
4	70%	30%	83.33%	82.35%	100%
5	65%	35%	79.37%	79.03%	100%
6	60%	40%	76.39%	79.10%	94.64%
7	55%	45%	72.84%	78.87%	88.89%
8	50%	50%	75.56%	80.00%	91.43%

Table 4.14 shows the results of tests performed with the Decision Tree Algorithm by applying PCA with the number of components C=6. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 72.84%, precision 78.87& and recall 88.89%, while the highest accuracy was found in split data 70%:30% with an accuracy value of 83.33%, precision 82.35%, and recall 100%.

4.4.6 Decision Tree Algorithm Test Results with PCA 7 Components (C=7)

The results of the Decision Tree classification experiment with the number of PCA components that are 7 (C = 7) are presented in Table 4.15.

Table 4.15 Decision Tree Algorithm Test Results with PCA (C=7)

Decision Tree - PCA 7 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	81.48%	80.77%	100%
2	80%	20%	75.00%	78.79%	92.86%
3	75%	25%	71.11%	77.50%	88.57%
4	70%	30%	83.33%	82.35%	100%
5	65%	35%	74.60%	77.97%	93.88%
6	60%	40%	76.39%	79.10%	94.64%
7	55%	45%	72.84%	78.87%	88.89%
8	50%	50%	75.56%	80.00%	91.43%

Table 4.15 shows the results of tests performed with the Decision Tree Algorithm by applying PCA with the number of components C=7. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 72.84%, precision 78.87& and recall 88.89%, while the highest accuracy was found in split data 70%:30% with an accuracy value of 83.33%, precision 82.35%, and recall 100%.

4.4.7 Decision Tree Algorithm Test Results with PCA 8 Components (C=8)

The results of the Decision Tree classification experiment with the number of PCA components which is 8 (C = 8) are presented in Table 4.16.

Table 4.16 Decision Tree Algorithm Test Results with PCA (C=8)

Decision Tree - PCA 8 Component					
No	Training data	Test Data	Accuracy	Precision	Recall
1	85%	15%	77.78%	77.78%	100%
2	80%	20%	75.00%	78.79%	92.86%
3	75%	25%	80.00%	79.55%	100%
4	70%	30%	83.33%	82.35%	100%
5	65%	35%	74.60%	77.97%	93.88%
6	60%	40%	76.39%	79.10%	94.64%
7	55%	45%	72.84%	78.87%	88.89%
8	50%	50%	75.56%	80.00%	91.43%

Table 4.16 shows the results of tests performed with the Decision Tree Algorithm by applying PCA with the number of components C=8. The experimental results of eight split data found that the lowest accuracy in split data was 55%:45% with an accuracy of 72.84%, precision 78.87& and recall 88.89%, while the highest accuracy was found in split data 70%:30% with an accuracy value of 83.33%, precision 82.35%, and recall 100%.

4.5 Discussion

The results showed that with the Naïve Bayes algorithm without PCA had the lowest accuracy of 33.33% in split data 85%:15%, while the highest accuracy with a value of 70.88% in split data 60%:40%. The results of the study by applying PCA and Naïve Bayes algorithm found the lowest accuracy with a value of 71.60% in split data 55%: 45% with the number of PCA components namely C = 4, while the highest accuracy with a value of 85.19% in split data 70% : 30% with the number of PCA components namely C = 6.

The results of research with the Decision Tree algorithm without PCA have the lowest accuracy of 67.90% in split data 55%:45%, while the highest accuracy with a value of 77.78% in split data 85%:15%. The results of the study by applying PCA and the Decision Tree algorithm found the lowest accuracy with a value of 71.60% in split data 55%: 45% with the number of PCA components namely C = 5, while the highest accuracy with a value of 83.33% in split data 70% : 30% with the number of PCA components namely C = 6. Based on the test results, the highest accuracy in this study was obtained in the Naïve Bayes algorithm with PCA with an accuracy value of 85.19%, split data 70%: 30%, the number of PCA components C = 6.

5. CONCLUSION AND ADVICE

5.1 Conclusion

In this study the use of the naïve bayes and Decision Tree methods is very influential on the use of PCA, this is shown by a significant difference in accuracy by using PCA on the Naive Bayes algorithm to get the highest value of 85.19% and the lowest value of 71.60%, without using PCA with the highest value of 70.83% and the lowest value of 33.33%. The same thing also happens to the Decisison Tree algorithm where the use of PCA greatly affects the accuracy value, namely the highest value with PCA 83.33% and the lowest value 71.60% while without PCA get the highest value 77.78% and the lowest value 67.97%. So with the results obtained, it can be tied that both algorithms are capable and reliable to solve existing problems.

6. REFERENCES

[1] Ardilla, Y., Manuhutu, A., Ahmad, N., Hasbi, I., Manuhutu, M. A., Ridwan, M., & Wardhani, A. K. (2021). *DATA MINING DAN APLIKASINYA*. Penerbit Widina.

[2] Azmi, B. N., Hermawan, A., & Avianto, D. (2023). Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *4*(4), 281–290. https://doi.org/10.35746/jtim.v4i4.298

[3] Dr. Abdul Kahar, M. P., & Dr. Rahmat Fadhli, E. M. (2021). *Beasiswa Pemutus Mata Rantai Kemiskinan*. Indonesia Emas Group.

[4] Mulyani, A., Kurniadi, D., Nashrulloh, M. R., Julianto, I. T., & Regita, M. (2022). The Preduction of PPA and KIP-Kuliah Scholarship Recipients Using Naive Bayes Algorithm. *Jurnal Teknik Informatika (JUTIF)*, *3*(4), 821–827.

[5] Rini, O., & Kunang, S. O. (2021). Implementasi Data Mining Menggunakan Metode Naive Bayes Untuk Penentuan Penerima Bantuan Program Indonesia Pintar (Pip) (Studi Kasus : Sd Negeri 9 Air Kumbang). *Bina Darma Conference on …*, 714–722.

[6] Tempola, F., Rosihan, R., & Adawiyah, R. (2021). Holdout Validation for Comparison Classfication Naïve Bayes and KNN of Recipient Kartu Indonesia Pintar. *IOP Conference Series: Materials Science and Engineering*, *1125*(1), 012041. https://doi.org/10.1088/1757-899x/1125/1/012041

[7] Utamajaya, J. N., Mentari, A., & Masnunah, S. (2019). Penerapan Algoritma Naive Bayes Untuk Penentuan

Calon Penerima Beasiswa PIP Pada SDN 023 Penajam. *Jurnal Sistem Informasi*, *3*(1), 11–17.

[8] Warella, S. Y., Revida, E., Abdillah, L. A., Pulungan, D. R., Purba, S., Firdaus, E., Tjiptadi, D. D., Faisal, M., Lie, D., & Butarbutar, M. (2021). *Penilaian Kinerja Sumber Daya Manusia*. Yayasan Kita Menulis.