RESEARCH ARTICLE OPEN ACCESS

Optimization of Algorithm C4.5, Naive Bayes With Particle Swarm Optimization in Predicting Career Suitability of Vocational High School Students: Case Study of SMKN 1 Rangkasbitung

Rini Marlina[1], Adrianto[2], Wahyu Amaldi[3], M.Johan Budiman[4],
[1234]Master of Computer Science Study Program, Postgraduate Program, Budi Luhur University
E-mail: [1] rinimarlinasmk@gmail.com, [2]4dr14nt0@gmail.com, [3]wahyu.amaldi@outlook.com, [4]mjbudiman@gmail.com

----------**********----------

Abstract:

"SMK Bisa"(=Vocational High School Can) be one of the tagline issued by the Ministry of Education and Culture in order to face the free market of ASEAN Economic Community(=MEA), so that the graduates of Vocational schools are ready to compete in the work field. Career is one aspect that cannot be separated from vocational school graduates, because one of the main goals of vocational school graduates is that they are absorbed in the work field or working. Some of the graduates who work are suitable with the majors in Vocational Schools but many of them are working not suitable with their majors. This study discusses the prediction of the suitability of Vocational students' careers. The method used in predicting career suitability uses C4.5 and Naive Bayes methods which are optimized using PSO. The results of the two methods used are increasing accuracy in both methods so that optimization produces algorithms that are superior between the two methods. From the superior algorithms that are used to predict the suitability of Vocational students' careers. After conducting the research, the accuracy of the C4.5 method is 89.14%, Naive bayes 86.86%. Furthermore, these two methods are optimized using PSO and the accuracy results is increasing. The result optimized accuracy of C4.5 becomes 93.80% while Naive Bayes is optimized to be 89.80%. Thus there is an increase in accuracy in both methods.

Key Words — **Career Suitability, Algorithm C4.5, Naive Bayes, PSO**

----------**********----------

I. PREFACE

To get good Human Resources (HR) in the company is by recruiting workers. Workers who have been recruited must be placed suitable with their competence in accordance with the rules of the right person in the right place. The existing process, often times, the competence or interest of students is not suitable with the expertise they have. Vocational High School is one of the secondary schools that preparethe studentsto work directly after graduated. But the ability of students during school with the needs in the work field is sometimes not the same so there is an unsuitable career of vocational students, the majors when in a vocational school with a career after graduating from vocational high school is not suitable as expected. For this reason, the writerconducts this research about the unsuitability career from vocational schools graduates using selected algorithms.

In predicting the career suitability of vocational high school students, a method is needed that can help minimize the impact of mistakes when choosing a career, namely by grouping data from data mining. Data mining is a discipline of knowledge that studies methods to extract knowledge or patterns from a data (*knowledge discovery in database*).[1]

Data mining can be used to group the data, predict, estimate, and determine association rules in an existing data. The need for data mining because

of the large amount of data that can be used to produce useful information and knowledge.

Data mining has a variety of methods, the method that the writer will use in this research is algorithm C4.5, Naive Bayes is optimized using Particle Swarm Optimization. The purpose of this study is to prove that the Algorithm C4.5, Naive Bayes optimized by PSO can develop better performance for vocational students' career recommendations.

Previous research discussing students' careers was discussed by Roshani and Deshmukh.[2] who predicted students career selection using the Naive Bayes method, K-Star and SVM accuracy of this study reached 89.6% for Naive Bayes, 89.2% for K-Star and 89.2% for SVM and there is an increase in accuracy after optimized using Adaboost. Takci et al.[3] discusses the measurement of career interests of high school students using the C4.5, SVM, Naive Bayes and MLP methods resulting in an average accuracy of 59.25%, while Shaymaa et al.[4] discusses the correlation of values with the characteristics of learningusing PLSA and LDA.

II. THEORETICAL BASIS

Data mining is a series of processes to explore added value in the form of information that has not been known manually from a database. The information generated is obtained by extracting and recognizing important or interesting patterns from the data contained in the database. Data mining is mainly used to search for knowledge contained in large databases so it is often called *Knowledge Discovery Databases* (KDD) (Tri Wulandari,Retno, 2017).[2]

2.1 Career Definition

Dillard (1985) distinguishes between job and career. According to him, job refers to work that does not continue and may be temporary. Therefore a job generally requires only a little expertise, little education, and a little dedication. While work as a career implies the existence of education and training, commitment, and is the path of work life chosen by individuals. [3]

2.2 Algorithm C4.5

This algorithm was developed to improve the ID3 algorithm. This algorithm is used in binary results as seen in CLS. So besides having features like ID3, C4.5 also has several different features that form the characteristics of ID3.

Calculate the roots of a tree. The root will be taken from the attribute that will be selected, by calculating the gain value of each attribute, the highest gainvalue which will be the first root.[4] Before calculating the gain value of an attribute, first calculate the entropy value. To calculate the entropy value the formula is used:

$$entropy\ (S) = \sum_{i=1}^{n} -pi\ \log_2 pi$$

Explanation :

S = Case set
n = number of partition S
P_i = proportion S_i toward S

Then calculate the gain value using the formula:

$$Gain\ (S, A) = entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si)$$

Explanation :

S = case set
A = Feature
n = number of partition attribute A
|Si| = Proportion Si towards S
|S| = number of case in S

2.3 Naive Bayes

Naive Bayes is a method used in statistics to calculate the probability of a hypothesis, Naive Bayes calculates the probability of a class based on the attributes possessed and determines the class that has the highest probability [5]

The general form of bayes theorem is as follows:

$$P(H|X) = P(X|H)\ P(H)\ P(X)\ (1)$$

Namely :

X = Data with unknown class
H = Hypothesis data X is a specific class
P(H|X) = Probability of hypothesis H based on condition X (posterior probability)
P(H) = Probability of hypothesis H (prior probability)

2.4 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a global optimization method introduced by Kennedy and Eberhart in 1995 based on research on the behavior of birds and fish flocks. [6] Each particle in Particle Swarm Optimization has the speed of particles moving in the search space with a dynamic speed adjusted to their historical behavior. Therefore, particles have a tendency to move towards better search areas during the search process. In the PSO algorithm there are several processes as follows:

1. **Initialization**
   a. Initializing the initial speed at the 0 iteration, it can be ascertained that the initial speed value of all particles is 0.
   b. Initializing the initial position of the particle In the 0 iteration, the initial position of the particle is generated by the equation:

$$x = xmin + [0,1] \times (xmax - xmin)$$

   c. *pBest*and *gBest*initialization At the 0 iteration, pBest will be equalized to the initial position of the particle. While gBest is chosen from one pBest with the highest fitness.

2. **Speed Update**
   To perform speed updates, use the following formula:

$$vi, t+1 = w.vi, t + c1.r1(Pbesti,jt - xi,jt) + c2.r2(Gbestg,jt - xi,jt)$$

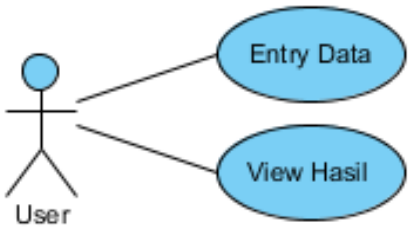
3. **Position Update and fitness calculation**
   To update the position, the following formula is used:

$$xi, t+1 = xi, t + vi, t+1$$

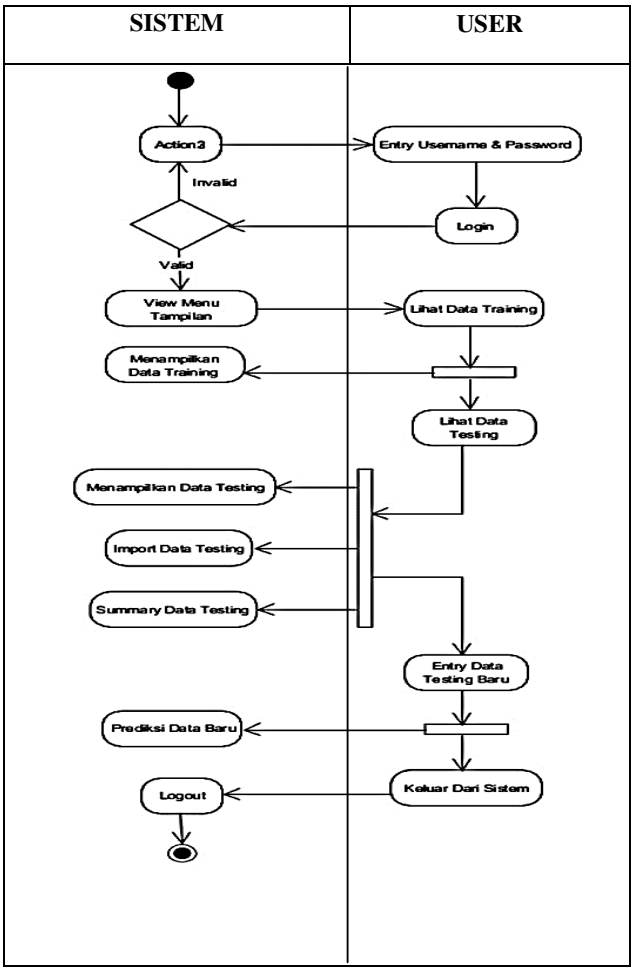
4. **Update pBest and gBest**
   Comparison between the pBest in the previous iteration with the results of the position update. Higher fitness will become a new pBest, the newest pBest that has the highest fitness value will be the new gBest.

III. SYSTEM DESIGN AND APPLICATION

Unified modelling language (UML) designfor this prototype is using use case diagram and activity diagram.



Picture III-1 Use Case



Picture III.2 Activiy Diagram

IV. RESULT AND DISCUSSION

Data collection was done by collecting data of alumni from SMKN 1 Rangkasbitung, graduating years in 2016 and 2017 for about 454 data. Theattribute data consists of 13 attributes. Wherein

12 predictor attributes and 1 result attribute. Of all the sampling data will be divided into training data and data testing as a model test through several stages of the model formation process, and new data as a model test.

Table IV-1
Attribute, Value and Description

No	Attribute	Value	Type	Description
1	**Gender**			
		Male	String	Students Gender
		Female		
2	**Majors**			
		MultiMedia	String	The choosen major
		Computer and networking technique		
		Accounting		
		Office administration		
		Marketing		
3	**Average score of Productive Subjects on 1st Semester**			
		Productive Score SM1>78	Real	Productive score of expertise on 1st semester
		Productive Score SM1<=78		
4	**Average score of Productive Subjects on 2nd Semester**			
		Productive Score SM2>82	Real	Productive score of expertise on 2nd semester
		Productive Score SM2<=82		
5	**Average score of Productive Subjects on 3rd Semester**			
		Productive Score SM3>71	Real	Productive score of expertise on 3rd semester
		Productive Score SM3<=71		
6	**Average score of Productive Subjects on 4th Semester**			
		Productive Score SM4>82	Real	Productive score of expertise on 4th semester
		Productive Score SM4<=82		
7	**Average score of Productive Subjects on 5th Semester**			
		Productive Score SM5>81	Real	Productive score of expertise on 5th semester
		Productive score SM5<=81		
8	**Industrial work practice score**			
		Apprentice >77	Real	Industrial work practice score
		Apprentice<=77		
9	**Expertise Competence Test Score**			
		ECT >83	Real	Expertise Competence Test Score
		ECT <=83		
10	**National Examination Score**			
		NE>220	Real	National Exam Score
		NE<=220		
11	**Graduation year**			
		2016	Integer	Students graduation year
		2017		
12	**Status**			
		Working	String	Students Status
		College		
13	**Career**			
		Suitable	String	Prediction of students Suitability Career
		Not Suitable		

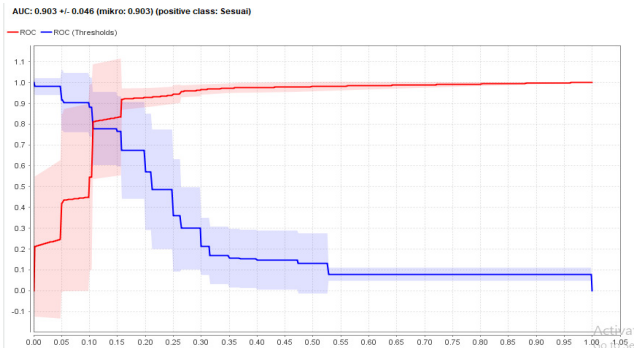
4.1 Testing Result of Algorithm C4.5

The test was carried out on all data training with confusion matrix consisting of accuracy, precision and recall carried out on data sets processed using algorithm C4.5, confusion matrix testing for datasets processed using model of algorithm C4.5. *Confusion Matrix* value from model algorithm C4.5

Table IV-2
*Confusion Matrix*value algorithm C4.5 with PSO

Prediction		**Not suitable**	**Suitable**	***Class Precision***
Accuracy 89,14%	Not suitable prediction	168	22	88.42%
	Suit-able prediction	27	234	89.66%
	Class Recall	86.15%	91.41%	
Precision 89,91%	Not suitable prediction	168	22	88.42%
	Suitable predicttion	27	234	89.66%
	Class Recall	86.15%	91.41%	
Recall 91,43%	Not suitable prediction	168	22	88.42%
	Suitable predicttion	27	234	89.66%
	Class Recall	86.15%	91.41%	

The following is the ROC curve generated by RapidMiner with the algorithm C4.5 model and produces an AUC value of 0.903.



Picture IV-1 ROC curve algorithm*C4.5*

4.2. Model Design ofAlgorithm PSO + C4.5

The quality of each attribute in the algorithm C4.5 has not shown a tendency to go to a certain value between 0 and 1. Therefore, giving the quality of each attribute is needed because not all attributes have an effect on the accuracy. By increasing the quality of the attribute, the resulting accuracy will increase. Particle Swarm Optimization (PSO) is used to helpselecting which attributes are influential so that it can improve the accuracy of the algorithm C4.5.

A. TestingResult Model Algorithm PSO + C4.5

The test was carried out on all data training with confusion matrix consisting of accuracy, precision and recall carried out on data sets processed using algorithm C4.5 on PSO using RapidMiner.

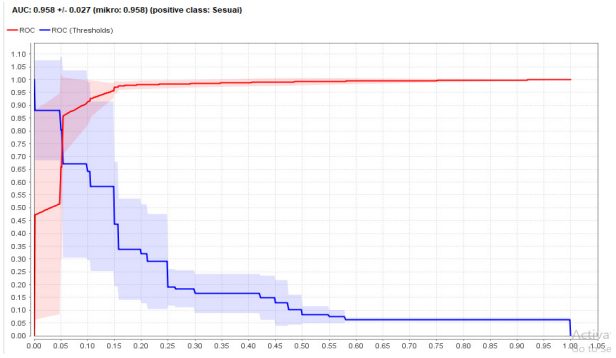
The accuracy level of the algorithm Particle Swarm Optimization (PSO) + algorithm C4.5 displayed is a final experiment by giving a value to the population size parameter of 30 and a maximum number of generation of 50. The results of Particle Swarm Optimization (PSO) + algorithm C4.5 are processed by RapidMiner tools displayed through Confusion Matrix tables and ROC Curves. The results are as follows:

Table IV-3
*Confusion Matrix*value algorithm C4.5 with PSO

Parameter PSO		Prediction		Execution time
1	Posize = 5 ; Generate = 30	Accuracy	93,56%	2 Second
		Precision	93,99%	
		Recall	94,95%	
		AUC	0,964	
2	Posize = 15 ; Generate = 30	Accuracy	93,82%	4 Second
		Precision	94,56%	
		Recall	94,95%	
		AUC	0,953	
3	Posize = 20 ; Generate = 30	Accuracy	93,58%	6 Second
		Precision	94,63%	
		Recall	94,18%	
		AUC	0,958	
4	Posize = 10 ; Generate = 40	Accuracy	93,35%	4 Second
		Precision	93,99%	
		Recall	94,52%	
		AUC	0,952	
5	Posize = 15 ; Generate = 40	Accuracy	93,80%	9 Second
		Precision	94,66%	
		Recall	94,52%	
		AUC	0,966	
6	Posize = 10 ; Generate = 50	Accuracy	93,80%	8 Second
		Precision	94,38%	
		Recall	94,92%	
		AUC	0,963	
7	Posize = 15 ; Generate = 50	Accuracy	93,80%	8 Second
		Precision	94,66%	
		Recall	94,52%	
		AUC	0.966	
8	Posize = 20 ; Generate = 50	Accuracy	93,81%	10 Second
		Precision	94,94%	
		Recall	94,54%	
		AUC	0.957	
9	Posize = 25 ; Generate = 50	Accuracy	94,00%	13 Second
		Precision	94,44%	
		Recall	95,32%	
		AUC	0.962	
10	Posize = 30 ; Generate = 50	Accuracy	93,80%	15 Second
		Precision	94,37%	
		Recall	94,95%	
		AUC	0.958	

B. ROC Curve algorithm PSO+C4.5

The following is the ROC curve generated by Rapidminer using the algorithm PSO + C4.5 and produces an AUC value 0.980.



Picture IV-2 ROC curve algorithmPSO+*C4.5*

4.3. Classification of Algorithm*Naive Bayes*

The Naïve Bayes algorithm calculation using training data in Table IV-13 begins with prior probability calculation to determine the suitable and not suitable values for all records. In the training data the number of records is 451, where the prediction of the students suitability career is 256 and the students with not suitable career is 195. Following the results of prior probability calculations are shown in the table below:

A. Algorithm*Naive Bayes*

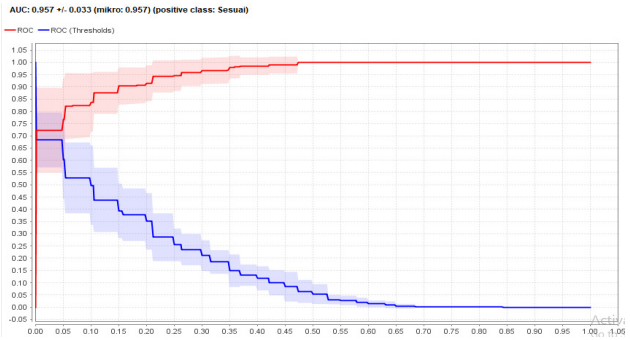
The test was conducted on all data training with confusion matrix consisting of accuracy, precision and recall carried out on data sets that were processed using the Naive Bayes algorithm using Rapidminer.

Table IV-4
*Confusion Matrix*value algorithm*Naive Bayes*

Prediction		Not Suitable	Suitable	Class Precision
Accuracy 86,48%	Not suitable prediction	179	45	79,91%
	Suitable prediction	16	211	92,95%
	Class Recall	91,79%	82,42%	
Precision 93,15%	Not suitable prediction	179	45	79,91%
	Suitable prediction	16	211	92,95%
	Class Recall	91,79%	82,42%	
Recall 82,45%	Not suitable prediction	179	45	79,91%
	Suitable prediction	16	211	92,95%
	Class Recall	91,79%	82,42%	

B. ROC Curve *Naive Bayes*on*RapidMinner*

The following is the ROC curve generated by Rapidminer using the Naive Bayes algorithm and produces an AUC value of 0.758.



Picture IV-3ROC curve algorithm *Naive Bayes*

4.4 Model Design of Algorithm PSO+*Naive Bayes*

The model proposed in the study is about the prediction of students' career suitability by applying Naïve Bayes based on Particle Swarm Optimization (PSO). The application of Naïve Bayes algorithm based the PSO refers to determining the right population size.

A. Testing Result Algorithm PSO+*Naive Bayes*

In this study, the model was tested by using 10 cross validation techniques, in which this process divides the data randomly into 10 parts. The testing process begins with the formation of a model with data in the first part. The model that is formed will be tested in the remaining 9 data sections. After that the accuracy process is calculated by seeing how much data that has been classified correctly.

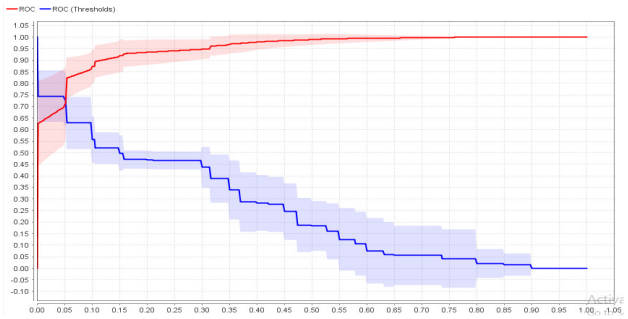
Table IV-5
Confusion Matrix PSO+*NaiveBayes*on*RapidMinner*

Parameter PSO		Prediction		Execution time
1	Posize = 5 ; Generate = 30	Accuracy	88,91%	1 Second
		Precision	92,32%	
		Recall	88,26%	
		AUC	0,951	
2	Posize = 15 ; Generate = 30	Accuracy	89,34%	2 Second
		Precision	91,74%	
		Recall	89,48%	
		AUC	0,947	
3	Posize = 20 ; Generate = 30	Accuracy	89,36%	3 Second
		Precision	91,83%	
		Recall	89,43%	
		AUC	0,949	
4	Posize = 10 ; Generate = 40	Accuracy	88,70%	2 Second
		Precision	92,12%	
		Recall	87,95%	
		AUC	0,956	

5	Posize = 15 ; Generate = 40	Accuracy	89,14%	4 Second
		Precision	91,99%	
		Recall	88,69%	
		AUC	0,953	
6	Posize = 10 ; Generate = 50	Accuracy	89,37%	3 Second
		Precision	91,96%	
		Recall	89,45%	
		AUC	0,950	
7	Posize = 15 ; Generate = 50	Accuracy	88,70%	5 Second
		Precision	94,17%	
		Recall	85,51%	
		AUC	0,962	
8	Posize = 20 ; Generate = 50	Accuracy	89,38%	6 Second
		Precision	91,88%	
		Recall	89,46%	
		AUC	0,936	
9	Posize = 25 ; Generate = 50	Accuracy	89,16%	8 Second
		Precision	86,66%	
		Recall	96,08%	
		AUC	0,939	
10	Posize = 30 ; Generate = 50	Accuracy	89,80%	9 Second
		Precision	93,90%	
		Recall	87,91%	
		AUC	0,954	

B. ROC Curve PSO + *Naive Bayes* on *RapidMinner*

The following is the ROC curve produced by Rapidminner using the PSO + Naive Bayes algorithm and produces an AUC value 0.954.

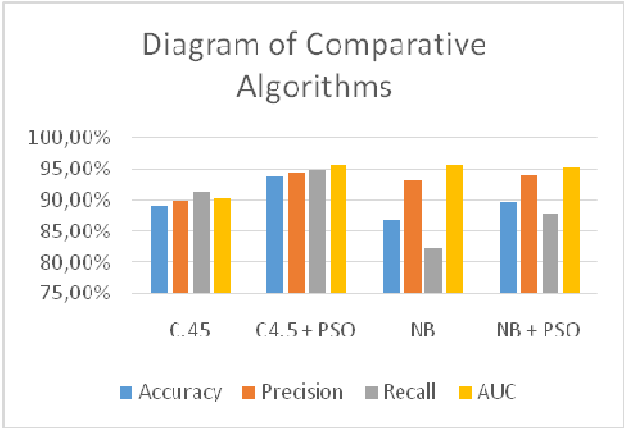


Picture IV-4 ROC curve algorithm*Naive Bayes+PSO*

4.5. Comparative Results

The model of C4.5 and Naive Bayes algorithmsto predict the career suitability of Vocational High School 1Rangkasbitung studentsin the level of accuracy produces a comparison of accuracy values and AUC. Using datasets of 2016 and 2017 alumni as test data. Algorithm C4.5 after optimized with PSO, get the highest accuracy value that is equal to 92.33%, while the Naive Bayes algorithm that is optimized with PSO gets an accuracy 91.33%.

Picture IV-5 Diagram Comparative Results of Algorithm Model C4.5, C4.5+PSO, *NB*, *NB*+PSO



V. CONCLUSION

Based on the research that the writer did in the optimization of C4.5 and Naïve Bayes algorithms which were optimized with PSO, it can be concluded as follows:

1. The implementation of the C4.5 and Naïve Bayes algorithms that are optimized using PSO both can increase the accuracy.
2. In the case study of Algorithm C4.5optimized with PSO get higher accuracy values compared to Naïve Bayes optimized with PSO in the case of vocational students' career predictions.
3. Algorithm C4.5 before optimization gets an accuracy value of 89.14% after optimized with PSO gets an accuracy value of 93.80%.
4. *Naive Bayes*before being optimized gets an accuracy value of 86.48% after being optimized with PSO gets an accuracy value of 89.80%.

Although algorithm C4.5 has high accuracy values, further research is needed, the following can be added to improve accuracy and performance, namely:

1. Using other classification algorithms contained in data mining, such as*Support Vector Machine*, *K-Nearest Neighbord.*

2. Using other optimizations such as *Adaptive Boost,Genetic Algorithm* to increase accuracy.

BIBLIOGRAPHY

[1] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*, Informatika,2017

[2] Roshani,Deshmukh," An incremental ensemble of classifiers as a technique for prediction of student's career choice",First International Conference on Network & soft Computing,IEEE,2014

[3] Takci et al,"Measurement of the appropriateness in career selection of the high school students by using data mining algorithms : A case study"Proceedings Of The Fedsis,Prague,2017

[4] Sorour et al,"Correlation of Grade Prediction Performance with Characteristics of Lesson Subject", 15th International Conference on Advanced Learning Technologies",IEEE,2015

[5] Wulandari Retno,' *Data mining dan Aplikasi RapidMiner*",Gava Media,2017

[6] I.Juwitaningrum,"*Program Bimbingan Karir untuk Meningkatkan Kematangan Karir Siswa SMK*",PSIKOPEDAGOGIA,Jurnal Bimbingan dan Konseling,Vol 2,No 2,2013

[7] R.H Pambudi,B.D Setiawan,Indriati,"*Penerapan Algoritma C4.5 dalam Program Untuk Memprediksi Kinerja Siswa Sekolah Menengah*"Jurnal Pengembangan Teknologi Informasi dan Komputer,Vol 2,No.7,Juli 2018

[8] M.S Akbar, Rochimah," *Prediksi Cacat Perangkat Lunak dengan Optimasi Naive Bayes Menggunakan Pemilihan Fitur Gain Ratio*" Jurnal Sistem dan Informatika,Vol 11,No 1,November 2016

[9] H.Muhammad,C.A Prasojo,N.Afifah, L.Surtiningsih, I.Cholissodin," *Optimasi Naive Bayes Classifier dengan Menggunakan Particle Swarm Optimization Pada Data Iris*",Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK),Vol 4, No 3,September 2017.