

Entrepreneurship Intention Prediction using Decision Tree and Support Vector Machine

Muhammad Dharma Tuah Putra Nasution¹, Andysah Putera Utama Siahaan², Yossie Rossanty¹, Solly Aryza²

¹Faculty of Social Science, Universitas Pembangunan Panca Budi, Medan, Indonesia

²Faculty of Science and Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia

dharma_nasution@dosen.pancabudi.ac.id, andiesiahaan@gmail.com,
yossie_rossanty@dosen.pancabudi.ac.id, sollyaryzalubis@gmail.com

ABSTRACT

This study discusses the prediction model of entrepreneurship intent for alumni. The data is obtained from the database of an online job market, alumni tracer and survey results to the alumni. This research applies the C4.5 decision tree algorithm to get a prediction model that shows the intention of entrepreneurship. Some essential indicators include Self-efficacy, Need for Achievement, Advisory Quotient, Locus of Control and Passion. The predictive model found that the best predictor was Self-efficacy which contributed to influence the entrepreneurship intention with a value of 79.7 percent. The authors recommend to educational institutions to foster candidate interest through curriculum improvement, field practice or learning models in and out of the classroom.

Keywords: *Prediction, Self-efficacy, Entrepreneurship, Passion, SVM*

1. INTRODUCTION

Information is the most important thing and is sent in all directions with the help of electrical appliance [1][2]. Large data should be categorized to be more focused [3][4]. Data mining is a concept used to discover hidden knowledge in a database. Data mining is a semi-automatic process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify potential and useful knowledge information stored in large databases [5][6]. Data mining is part of Knowledge Discovery in Databases process, consisting of stages such as data selection, pre-processing, transformation, data mining, and evaluation of results [7][8][9]. Data mining techniques can be broadly divided into two groups: verification and discovery. Verification methods generally include statistical techniques such as goodness of fit, and analysis of variance. Further discovery methods can be divided into predictive models and descriptive models [10].

Predictive techniques predict data by using known results from different data [11]. This model can be created based on the use of other historical data. The descriptive model aims to identify patterns or relationships between data and provide a way to explore the characteristics of the data being investigated [12]. Nowadays it is still a central issue of the university world of predictor factors and techniques used to predict the desire of graduates to start a new business after completing their college. Until now rarely found predictors and techniques that are reliable enough and accurate in predicting the level of desire graduates or alumni to start a

business, whether they will start a business or find a job or continue study to post-graduate. In this research, we will examine how the behavior of graduates to start a business that will use decision tree techniques to analyze and build a prediction model based on the behavior of college graduates and using the Support Vector Machine technique to classify graduates based on the predictive model obtained by the decision tree. Both methods are selected because the decision tree method is quite simple and widely used by other researchers in developing a model. Support Vector Machine (SVM) method is a relatively new technique in pattern recognition and is state of the art in recognition pattern and machine learning because of its reliability in processing many dimension data. This study takes the college area as one of the research domains in the field of data mining with the data source from university academic database and surveying 1300 graduates from several study programs. The research is expected to contribute to Universitas Pembangunan Panca Budi.

2. THEORIES

2.1 Data Mining

Data mining is an analysis of the dataset review to find unexpected relationships and summarize data in different ways than ever before, which can be understood and useful for data owners [13]. Data mining brings together techniques from machine learning, pattern recognition, statistics, databases, and visualizations to handling information retrieval problems from large databases [14]. The relationship sought in data mining can be a relationship between two or more in one dimension [15]. The term data mining and Knowledge Discovery in Database (KDD) are often used interchangeably to explain the process of extracting hidden information in an extensive database. The two terms have different concepts but are related to each other.

2.2 Algorithm C 4.5

Algorithm C 4.5 is one method to make decision tree based on training data that has been provided. Algorithm C 4.5 is the development of ID3. Some of the development done on C4.5 is as among others can overcome the missing value, can overcome continue data, and pruning. It is a well-known method of classification and prediction. The decision tree method transforms a huge fact into a decision tree that represents the rule. Rules can be easily understood with natural language [16]. It can also be expressed in the form of database languages such as Structured Query Language to search for records in certain categories. Decision trees are also useful for exploring data, finding the hidden relationship between some potential input variables with a target variable [17]. Because decision trees combine data exploration and modeling, decision trees are excellent as a first step in the modeling process even when used as final models of some other techniques. A decision tree is a structure that can be used to divide large datasets into smaller record sets by applying a set of decision rules. With each set of divisions, members of the result set become similar to each other [18].

A decision tree model consists of a set of rules to divide the number of heterogeneous populations into smaller, more similar ones by taking into account its destination variables. A decision tree may be built carefully manually, or it can grow automatically by applying one or more decision tree algorithms to model unclassified data sets. Destination variables are usually grouped with definite and decision tree models more lead to the probability calculation of each

record against these categories or to classify records by grouping them in one class. Decision trees can also be used to estimate the value of the continue variable although there are some techniques more appropriate for this case. Many algorithms can be used in decision tree formation, such as ID3, CART, and C4.5 [13]. The data in the decision tree is usually expressed in tabular form with attributes and records. The attribute states a parameter created as a criterion in the formation of a tree. One of the attributes is an attribute that states the data solution per data item called the target attribute. The attribute has values named with the instance.

2.3 Support Vector Machine (SVM)

Pattern Recognition is one of the fields in computer science, which maps data into specific predefined concepts [19]. This particular concept is called a class or category[3]. The application of pattern recognition is comprehensive, including recognizing the sound in the security system, reading the letters in the OCR, classifying the disease based on the diagnosis of the patient's medical condition and so on [20]. Various methods are known in pattern recognition, such as linear discrimination analysis, hidden Markov models to artificial intelligence methods such as artificial neural networks. One method that has recently received much attention as state of the art in pattern recognition is the Support Vector Machine (SVM) [21]. The fundamental principle of SVM is a linear classifier and further developed to work on nonlinear problems by incorporating the kernel trick concept in high-dimensional workspaces [22][23]. This development provides a stimulus of research interest in the field of pattern recognition to investigate the potential of SVM capability theoretically as well as regarding application [24].

3. METHODOLOGY

3.1 Data Collection Procedures

The authors collected data from developing a questionnaire to measure the correct linkage of data from alumni to the first dataset. The question is a 5-point Likert scale that is composed of "strongly agree" to "strongly disagree." The first dataset can be seen in the following table.

Table 1. First dataset

No.	Role	Name	Type
1	Label	Entrepreneurial Intention Study	Nominal
2	Reguler	Advisory Quotient	Nominal
3	Reguler	Self-efficacy	Nominal
4	Reguler	Locus of Control	Nominal
5	Reguler	Need for Achivement	Nominal
6	Reguler	Passion	Nominal

Table 1 describes the Entrepreneurial Intention attribute as a label that is the goal of the Advisory Quotient attributes, Self-efficacy, Locus of Control, Need for Achievement, and Passion.

3.2 Data Validity and Reliability

The accuracy used refers to whether the instruments used consistently measure each time and population. The surveys in this study were tested in a reliable internal timeframe and size that had a link between the test section. It is to ensure that the measurement of the instrument is done accurately. The internal consistency measurements in the survey use Cronbach's Alpha. Cronbach's Alpha at or above 0.60 is accepted as evidence of internal reliability [25]. The validity and reliability of the questionnaire are shown in the following table.

Table 2. Data reliability statistics

No	Variable	Cronbach's alpha	Amount
1	Self-efficacy	0.669	75
2	Need for Achievement	0.655	75
3	Advisory Quotient	0.663	75

3.3 Decision Tree Model

The following table is the proposed decision tree model.

Table 3. Decision tree model

Tree									
SE	high								
	NA	high							
		AQ	high						
			LC	high					
				PA	high				
					Locus of Control	> 24.500:	y(y=16, n=0)		
					Locus of Control	≤ 24.500			
						Advisory Quotient	> 23.500		
							Self-efficacy	> 23.500	y(y=8, n=0)
							Self-efficacy	≤ 23.500	y(y=7, n=1)
						Advisory Quotient	≤ 23.500:	y(y=11, n=1)	
			LC=	low					
				PA=	Low:	y(y=12, n=1)			
			LC=	medium					
				PA=	medium				
					Self-efficacy	> 23.500	y(y=12, n=0)		
					Self-	≤ 23.500			

					efficacy				
						Locus of Control	> 21.500	y(y=25, n=0)	
						Locus of Control	≤ 21.500		
							Advisory Quotient	> 23.500	y(y=11, n=1)
							Advisory Quotient	≤ 23.500	y(y=13, n=1)
		AQ=	Low						
			LC=	high:	y(y=34, n=0)				
			LC=	low					
				PA=	Low				
					Locus of Control	> 18:	y(y=6, n=0)		
					Locus of Control	≤ 18			
						Advisory Quotient	> 18.500:	y(y=15, n=0)	
						Advisory Quotient	≤ 18.500:	y(y=5, n=1)	
			LC=	medium					
				PA=	medium				
					Locus of Control	>21			
						Advisory Quotient	> 17.500:	y(y=5, n=1)	
						Advisory Quotient	≤ 17.500:	y(y=14, n=0)	
					Locus of Control	≤ 21:	y(y=5, n=1)		
		AQ=	medium						
			LC=	High					
				PA	high				
					Self-efficacy	> 23.500			
						Advisory Quotient	> 20.500		
							Advisory Quotient	> 21.500:	y(y=11, n=1)
							Advisory Quotient	≤ 24.500:	y(y=8, n=0)
						Advisory Quotient	> 17.500:	y(y=15, n=0)	
					Self-efficacy	≤ 23.500	y(y=14, n=0)		

3.4 Support Vector Machine Model

The graph of support vector machine model that will be used is presented in the following picture.

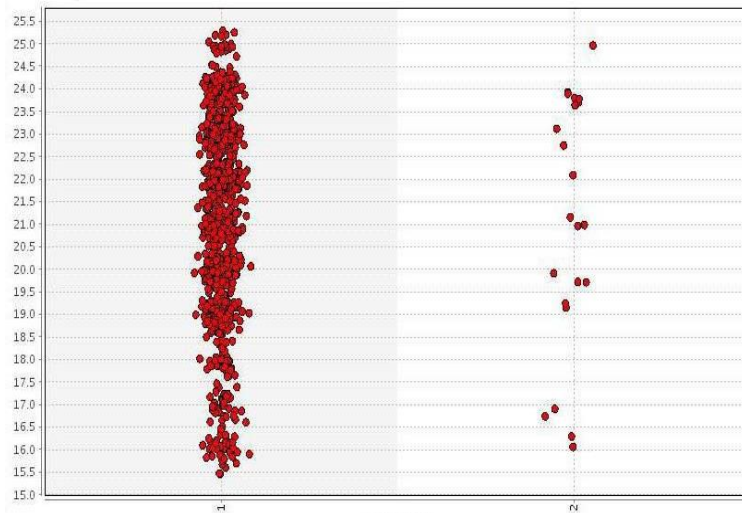


Figure 1. Graph model of SVM

Figure 1 explains that after the experiment using the support vector machine that each variable is in the data directly compared to the label predicate in the intention so that after translated into simple language is: Number of variables Advisory Quotient proxies into next predicate 1 and 2 which means 1 = yes, 2 = no.

4. RESULT AND DISCUSSION

In this experiment, the authors tested the significant reliability of data from the questionnaires distributed to alumni by using Cronbach's Alpha method for the sample data consisting of 75 data, while for the overall data as the data testing the author uses as many as 1254 data.

4.1 Descriptive Data Experiment Result

Data processing Descriptive statistics are used to present descriptive statistical data on a variable such as the amount of data, the size of the spread of the range of range, minimum, maximum, mean, standard deviation, and variance of the results of the overall data test can be seen in the following table.

Table 4. Descriptive Statistics of Testing Data

	N	Range	Min	Max	Mean	Std. Deviation	Variance
p1	1254	2	3	5	4,38	,657	,432

P2	1254	2	3	5	3,67	,731	,534
p3	1254	2	3	5	4,03	,824	,680
p4	1254	2	3	5	4,40	,633	,401
p5	1254	2	3	5	4,66	,507	,257
Self-efficacy	1254	7	13	20	16,76	1,788	3,196
p6	1254	2	3	5	4,37	,657	,432
p7	1254	2	3	5	3,66	,722	,521
p8	1254	2	3	5	4,02	,827	,685
p9	1254	2	3	5	4,39	,633	,401
p10	1254	2	3	5	4,65	,510	,261
Need for Achievement	1254	9	16	25	21,08	2,156	4,648
p11	1254	2	3	5	4,39	,659	,435
P12	1254	2	3	5	3,68	,743	,552
p13	1254	2	3	5	4,02	,823	,678
p14	1254	2	3	5	4,40	,638	,407
p15	1254	2	3	5	4,67	,504	,255
Advisory Quotient	1254	9	16	25	21,16	2,162	4,675
Valid (listwise)	N734						

Table 4 describes the function of descriptive data analysis is to provide an overview of the data that has been obtained. This general description can be a reference to see the characteristics of the data we get. To begin this analysis, the author performs descriptive analysis before performing any other analysis on the data. It is essential because with the descriptive analysis we can quickly correct the data we have already entered.

4.2 Data Frequency Experiment Result

4.2.1 Frequency Statistics of Self-efficacy Factor

Data processing frequency statistics of economic factors are used to display and describe data about economic factors of respondent choice. The results of the overall data test can be seen in the following table.

Table 5. Frequency Statistics of Self-efficacy Factors

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16	60	4,8	4,8	4,8
	17	60	4,8	4,8	9,5
	18	198	15,8	15,8	25,3
	19	258	20,6	20,6	45,9
	20	176	14,0	14,0	59,9
	21	234	18,6	18,6	78,5
	22	174	13,9	13,9	92,5

	23	63	5,0	5,0	97,5
	24	16	1,3	1,3	98,8
	25	15	1,2	1,2	100,0
	Total	1254	100,0	100,0	

Table 5 describes the frequency of data, in which the number of respondents who have data based on a particular Self-efficacy. Such as respondents with a total value of 16 there are 60 people, respondents with a value of 17 there are 60 people and so on. The percent value shows the percentage of the amount of data that has the Self-efficacy factor.

4.2.2 Frequency Statistics of Need for Achievement Factor

The Need For Achievement statistical data frequency data processing is used to display and describe data about the Need for Achievement factor of the respondent's choice. The results of the overall data test can be seen in the following table.

Table 6. Frequency Statistics of Need for Achievement Factor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16	41	3	3,3	3,3
	17	41	3,3	3,3	6,5
	18	43	3,4	3,4	9,9
	19	186	14,9	14,9	24,8
	20	193	15,4	15,4	40,2
	21	188	15,0	15,0	55,2
	22	166	13,2	13,2	68,4
	23	210	16,8	16,8	85,1
	24	166	13,2	13,2	98,4
	25	20	1,6	1,6	100,0
	Total	1254	100,0	100,0	

Table 6 describes the frequency, indicating the number of respondents who have data based on a specific Need for Achievement. For example, the respondent with a total score of 16 has 41 people, the respondent with the value of 17 there are 41 people and so on. The percent value shows the percentage of the amount of data that has the Need for Achievement factor.

4.2.3 Frequency Statistics of Advisory Quotient Factor

The statistical frequency data processing of the alumni learning facilities factor is used to display and describe the data about the factors of the alumni learning facilities from the respondents' choice. The results of the overall data test can be seen in the following table.

Table 7. Frequency Statistics of Advisory Quotient Factor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16	41	3,3	3,3	3,3
	17	41	3,3	3,3	6,5
	18	41	3,3	3,3	9,8
	19	166	13,2	13,2	23,0
	20	188	15,0	15,0	38,0
	21	190	15,1	15,1	53,1
	22	190	15,1	15,1	68,3
	23	188	15,0	15,0	83,2
	24	188	15,0	15,0	98,2
	25	21	1,8	1,8	100,0
	Total	1254	100,0	100,0	

Table 7 describes the frequency, indicating the number of respondents who have data based on certain Advisory Quotient. For example, the respondent with a total score of 16 has 41 people, the respondent with the value of 17 there are 41 people and so on. The percent value shows the percentage of the amount of data that has the Advisory Quotient factor while the frequency test for statistics, all data entered into the test data.

4.3 Decision Tree Experiment Result

It can be explained that the factor of Self-Efficacy has the most influence in research data of this research. The experiment shows that Self-Efficacy variable is at the topmost node, then compared with the mean value from Self-Efficacy factor to get advanced predicate. If the rule graph is translated into a simple language rule can be seen in the following table.

Table 8. Decision Tree result

Rule	Rule Statement	Predicate
1	If Self-efficacy = High	Yes
2	If Self-efficacy = High & Amount of Need for Achievement = High	Yes
3	If Self-efficacy = High & Amount of Advisory Quotient = High	Yes
4	If Self-efficacy = High & Amount of Locus of Control = High	Yes
5	If Self-efficacy = High & Amount of Passion = High	Yes

4.4 Decision Tree Validation

Decision tree validation is used to see the accuracy of the model of the behavioral behavior of alumni using the C4.5 algorithm in determining interest for entrepreneurship. In validating the rule decision tree, the author uses the split validation utility used to divide the two training areas and testing the data. In testing the data, the author uses utility apply to model and% performance. The design model validation can be seen in the following figure.

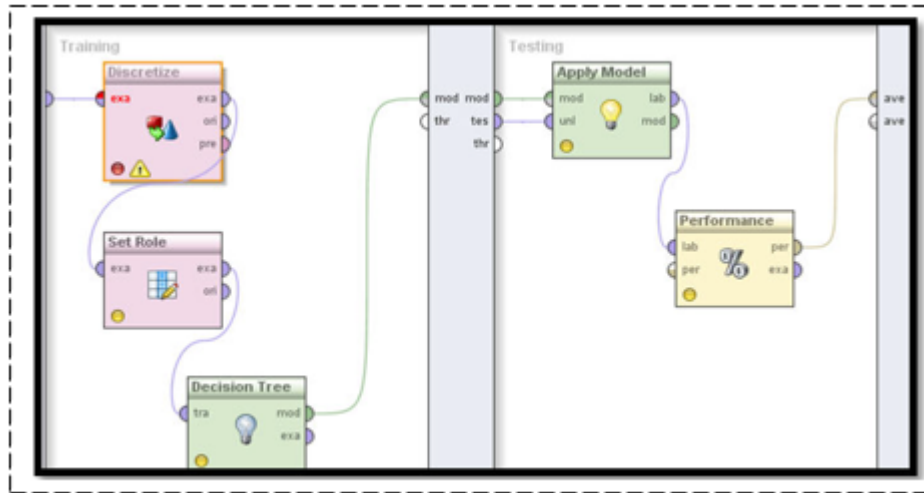


Figure 2. Decision Tree model

Figure 2 describes the training menu consisting of Discretize. It works to change the numeric value to nominal, then to create a label using Set Role and to get the author's roles using decision tree algorithm while author testing menu uses apply model and performance.

4.5 Support Vector Machine Experiment Result

To help see the dependence of the relationship of alumni data with questionnaires data can be seen the results of multiple scatter, predicate image with the value-related performance of five predictor variables consisting of Self-efficacy, Need for Achievement, Advisory Quotient, Locus of Control, and Passion.

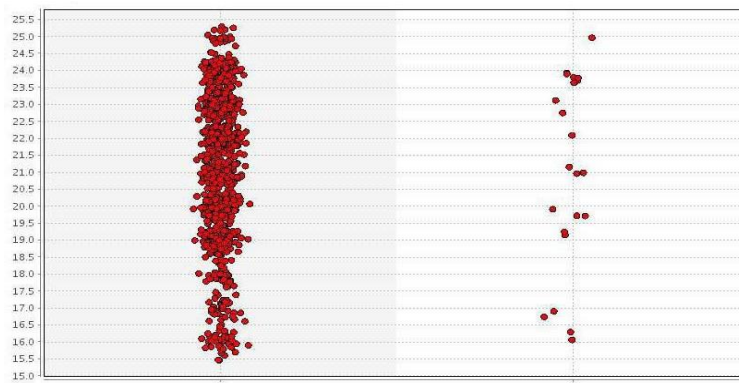


Figure 3. Relationship between Predicate and Passion

Figure 3 is the result of the data grouping of the Passion variable associated with the predicate.

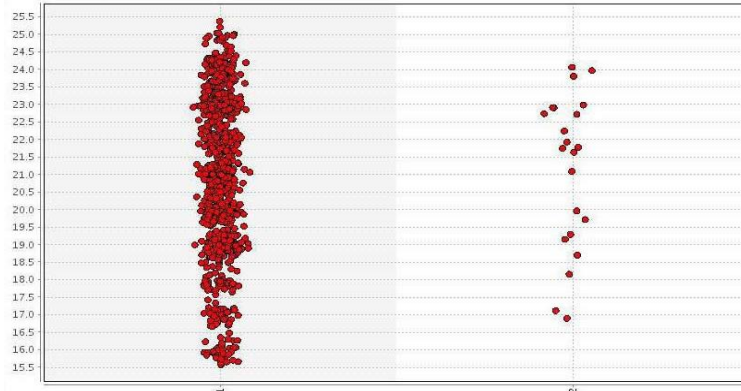


Figure 4. Relationship between Predicate with Need for Achievement

Figure 4 is the result of a grouping of data from the Need for Achievement variable associated with the predicate.

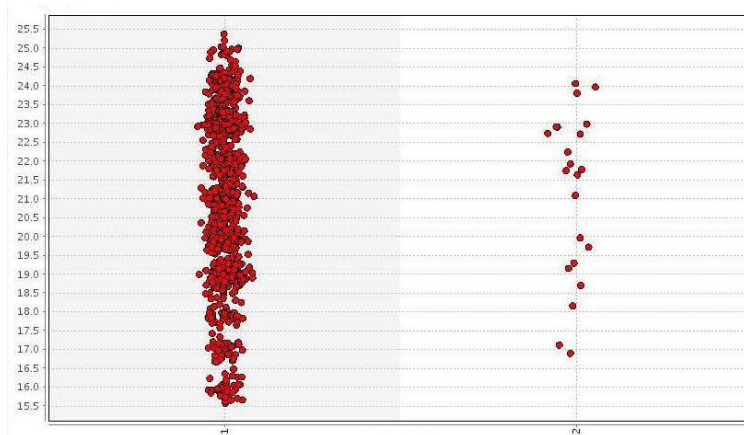


Figure 5. Relationship between Predicate and Self-Confidence

Figure 5 is the result of the data grouping of the Self-efficacy variable associated with the predicate.

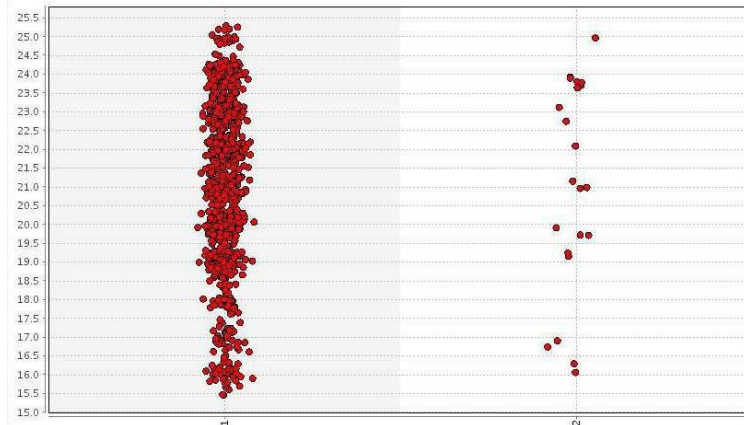


Figure 6. Relationship between Predicate and Locus of Control

Figure 6 is the result of the data grouping of the Locus of Control variable associated with the predicate.

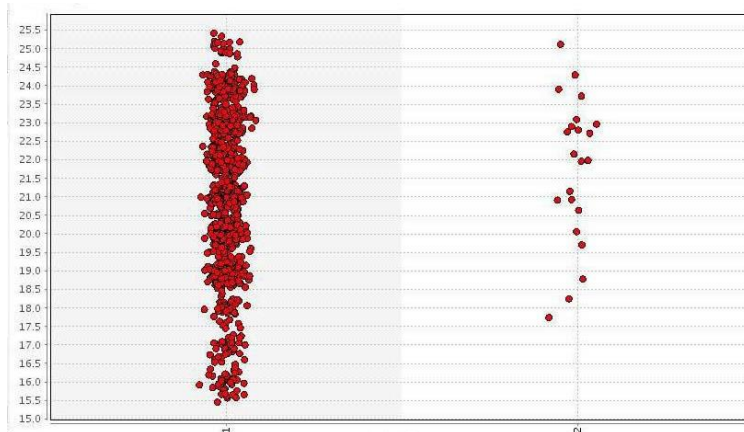


Figure 7. Relationship between Predicate and Advisory Quotient

Figure 7 is the result of the data grouping of the Advisory Quotient variable associated with the predicate. Experiments using the support vector machine state that each of the variables in the direct data is compared to the label predicate in question. The results of the experiment stated that:

- The number of Passion variables is projected into the predicate Next 1 and 2 which means 1 = yes, 2 = no.
- The number of Need for Achievement variables is projected into predicate Next 1 and 2 which means 1 = yes, 2 = no.
- The number of Self-efficacy variables is proportional to the predicate Continue 1 and 2 which means 1 = yes, 2 = no.
- The number of Locus of Control variables is projected into predicates 1 and 2 which means 1 = yes, 2 = no.

- The number of Advisory Quotient variables is projected into predicate 1 and 2 which means 1 = yes, 2 = no.

5. CONCLUSION

There is a model rule that can show the analysis and prediction of alumni intentions for entrepreneurship. In a case study on alumni of Panca Budi Development University that most have a predicate for entrepreneurship if the average factor Self-efficacy = High and Need for Achievement = High and Advisory Quotient = High. Algorithm C 4.5 is still considered as an algorithm that is very helpful in classifying data because the characteristics of classification data obtained clearly both in the form of tree decision structure (tree) and the rules induction rule if then, making it easier for users in digging information on the concerned. This study has shown that three of the five variables have a very significant correlation to the model rules of analysis and prediction on interest for self-entrepreneurship as proposed (Self-efficacy, Need for Achievement Advisory Quotient) with the predicate of entrepreneurial intention. All three predictors of the three variables contributed 82.8%, and the best variables of the Self-efficacy predictors contributed 79.8%.

REFERENCES

- [1] S. Aryza, M. Irwanto, Z. Lubis, A. P. U. Siahaan, R. Rahim, and M. Furqan, "A Novelty Design of Minimization of Electrical Losses in A Vector Controlled Induction Machine Drive," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 300, no. 1.
- [2] A. Putera, U. Siahaan, and R. Rahim, "Dynamic Key Matrix of Hill Cipher Using Genetic Algorithm," *Int. J. Secur. Its Appl.*, vol. 10, no. 8, pp. 173–180, Aug. 2016.
- [3] R. Rahim *et al.*, "Searching Process with Raita Algorithm and its Application," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, p. 012004, Apr. 2018.
- [4] S. Haryati, A. Ikhwan, D. Arisandi, A. Putera, and U. Siahaan, "Quality Assurance in Knowledge Data Warehouse," *Int. J. Sci. Res. Sci. Technol.*, vol. 3, no. 6, pp. 239–242, 2017.
- [5] L. Marlina, Muslim, and A. P. U. Siahaan, "Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)," *International J. Eng. Trends Technol.*, vol. 38, no. 7, pp. 380–383, 2016.
- [6] E. Turban, J. E. Aronson, and T. Liang, *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi, 2005.
- [7] W. Fitriani and A. P. U. Siahaan, "Comparison Between WEKA and Salford System in Data Mining Software," *Int. J. Mob. Comput. Appl.*, vol. 3, no. 4, pp. 1–4, 2016.
- [8] O. Maimon and M. Last, *Knowledge Discovery and Data Mining, The Info-Fuzzy Network (IFN) Methodology*. 2001.
- [9] R. Rahim, Mesran, A. Putera, U. Siahaan, and S. Aryza, "Composite performance index for student admission," *Int. J. Res. Sci. Eng.*, vol. 3, no. 3, pp. 68–74, 2017.
- [10] M. D. L. Siahaan, Elviwani, A. B. Surbakti, A. H. Lubis, and A. P. U. Siahaan, "Implementation of Simple Additive Weighting Algorithm in Particular Instance," *Int. J. Sci. Res. Sci. Technol.*, vol. 3, no. 6, pp. 442–447, 2017.
- [11] M. D. T. P. Nasution and Y. Rossanty, "Country of Origin as a Moderator of Halal Label and Purchase Behavior," *J. Bus. Retail Manag. Res.*, vol. 12, no. 2, pp. 194–201, 2018.

- [12] M. H. Dunham, *Data Mining Introductory and Advanced Topics*. New Jersey: Prentice Hall, 2003.
- [13] T. Larose D, *Data Mining Methods and Models*. New Jersey: Jhon Wiley & Sons, Inc., 2006.
- [14] A. Zanasi, C. A. Brebbia, and N. Ebecken, *Data Mining III*. WIT Press, 2002.
- [15] Supiyandi, M. I. Perangin-angin, A. H. Lubis, A. Ikhwan, Mesran, and A. P. U. Siahaan, "Association Rules Analysis on FP-Growth Method in Predicting Sales," *Int. J. Recent Trends Eng. Res.*, vol. 3, no. 10, pp. 58–65, 2017.
- [16] M. P. Sari, A. H. Mirza, and Fatmasari, "Decision Support System to Determine The Loan Nominal at KUD Mupakat Jaya Using Decision Tree Method," in *The 4th ICIBA 2015, International Conference on Information Technology and Engineering Application*, 2015, pp. 250–254.
- [17] E. Hajizadeh, H. D. Ardakani, and J. Shahrabi, "Application of Data Mining Techniques in Stock Markets: A survey," *J. Econ. Int. Financ.*, vol. 2, no. 27, pp. 109–118, 2010.
- [18] M. J. A. Berry and L. G. S., *Data Mining Techniques For Marketing, sales, Customer Relationship Management, Second Edition*,. Wiley Publishing, Inc., 2004.
- [19] R. Meiyanti, A. Subandi, N. Fuqara, M. A. Budiman, and A. P. U. Siahaan, "The recognition of female voice based on voice registers in singing techniques in real-time using hankel transform method and macdonald function," *J. Phys. Conf. Ser.*, vol. 978, no. 1, p. 012051, Mar. 2018.
- [20] R. Rahim *et al.*, "Combination Base64 Algorithm and EOF Technique for Steganography," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, p. 012003, Apr. 2018.
- [21] Chun-Fu Lin and Sheng-De Wang, "Fuzzy Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 464–471, Mar. 2002.
- [22] Kwang In Kim, Keechul Jung, Se Hyun Park, and Hang Joon Kim, "Support Vector Machines for Texture Classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1542–1550, Nov. 2002.
- [23] K. Huang, H. Jiang, and X.-Y. Zhang, "Field Support Vector Machines," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 1, no. 6, pp. 454–463, Dec. 2017.
- [24] N. Mayasari, "Comparison of Support Vector Machine and Decision Tree in Predicting On-Time Graduation (Case Study : Universitas Pembangunan Panca Budi)," *Int. J. Recent Trends Eng. Res.*, vol. 2, no. 12, pp. 140–151, 2016.
- [25] M. M. and J. J., *Research Design Explained*. New York: Holt, Rinehart and Winston, 1999.