International Research Journal of Computer Science (IRJCS)
Issue 01, Volume 5 (January 2018)
ISSN: 2393-9842
www.irjcs.com

A K-NEAREST ALGORITHM BASED APPLICATION TO PREDICT SNMPTN ACCEPTANCE FOR HIGH SCHOOL STUDENTS IN INDONESIA

Adi Tri Wibowo
Department of Informatics, Faculty of Computer Science
Universitas Mercu Buana, Indonesia
aditriwibowo@hotmail.com;
Devi Fitrianah
Department of Informatics, Faculty of Computer Science
Universitas Mercu Buana, Indonesia
devi.fitrianah@mercubuana.ac.id;

Manuscript History
Number: **IRJCS/RS/Vol.05/Issue01/JACS10083**
DOI: 10.26562/IRJCS.2018.JACS10083
Received: 05, December 2017
Final Correction: 24, December 2017
Final Accepted: 06, January 2018
Published: January 2018
Citation: Wibowo & Fitrianah (2017). A K-NEAREST ALGORITHM BASED APPLICATION TO PREDICT SNMPTN ACCEPTANCE FOR HIGH SCHOOL STUDENTS IN INDONESIA. IRJCS:: International Research Journal of Computer Science, Volume V, 09-20. **doi: 10.26562/IRJCS.2017.JACS10083**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract — Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) is one of the acceptances to enrol public universities which held simultaneously throughout Indonesia. In Indonesia, there is no application that can assist high schools in predicting the student acceptance to public universities yet. In this study, we propose an application that can predict SNMPTN results based on learning data. This prediction application will identify data-driven based on the SMAN 8 Jakarta alumni's report cards accepted and not accepted in SNMPTN. The data is from the average scores within semesters 1 until 5 of alumni data as training data in the classification process. We compare the training data with two testing data models (training data itself and 10-folds cross validation) as testing data. The algorithm that is used for this study is K-Nearest Algorithm. The result shows that the optimal parameters to gain good prediction only involve 5 parameters, they are the average score of semester 1, the average score of semester 2, the average score of semester 3, the average score of semester 4, and the average score of semester 5. The results show good accuracy, 80% for evaluating the science majoring alumni's data itself with K=3 and 89% for evaluating the social science majoring alumni's data itself with K=3. The prediction is then applied in a web based application which is developed utilizing the Relational Unified Process Framework. From this study we also find out that there are 5 out of 8 parameters that can be used in the SNMPTN prediction, they are average score in semester 1 until semester 5.
Keywords— Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN); Public University; Predicting; Classification; K-nearest Neighbour Algorithm;

I. INTRODUCTION

SNMPTN is an acronym of "*Seleksi Nasional Masuk Perguruan Tinggi Negeri*". SNMPTN is one of the acceptances to enroll public universities which held simultaneously throughout Indonesia. Unlike the other acceptance, SNMPTN is 'the invitation' and does not require a written test or exam anymore. SNMPTN has the greatest chance for acceptance. SNMPTN is a selection of potential students to enter public universities in the country level with the acceptance selection criteria based on the report cards. Since 2011, each year only about 20% of the total

applicants who received through SNMPTN from all over Indonesia, this reason that makes SNMPTN become one and only public universities acceptance which very tight and prestigious. Although SNMPTN has the biggest quota among the selection of public university acceptances, but SNMPTN is not easy as well. The intense competition often makes students feel pessimistic to register SNMPTN [1].

SNMPTN prediction is still manual which unaccountable until now. The prediction itself can be performed in many ways. One of the prediction method that can be utilized is the classification techniques [2]. One of the best classification techniques to use in data classification is the K-Nearest Neighbor (KNN) [3]. There is a research in prediction of SNMPTN [1] have been done utilized the average scores of semesters 3, 4, 5, 6. However, the correct data for SNMPTN indicator according to data from SNMPTN official website is the average scores of semesters 1, 2, 3, 4, and 5 [4]. In Indonesia, there is no application that can assist high schools in predicting the student entrance to public universities until now. Therefore, based on the above explanation, we propose a web-based application that is required to help the school in predicting their students in public universities through SNMPTN. This application is based on their alumni's data, which then based on previous data, we can see who is accepted and not accepted by public universities through SNMPTN based on their report cards. This prediction application will identify data-driven based on the SMAN 8 Jakarta alumni's report cards accepted and not accepted SNMPTN. The data used are SMAN 8 Jakarta alumni's data within 5 years (from 2013 - 2017). The weighted scores are calculated from the average scores from semesters 1 until 5. In this study, we used K-Nearest Neighbor as an algorithm to classify data. Last of all, our proposed prediction application will recommend the list of universities to the users to apply for SNMPTN acceptance which is beneficial for school to predict their student's acceptance in public universities.

II. LITERATURE REVIEW

A. K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a more flexible technique and the simplest machine learning algorithms because it is able to classify test data into label classes by searching for train data which is relatively the same as test data [1][5]. KNN is known to be one of the best state of the art classifiers and could achieve best accuracy [6][7][8]. K-nearest neighbors is an algorithm that falls under the category of supervised learning algorithms [9]. The classification as per this algorithm is done based on the distances between the training data and the testing data. The distances are calculated using distances such as the Euclidean Distance [10]. Based on the similarity between the training and the testing data the nearest k neighbors are selected. Here k is a positive integer. The label associated with these neighbors is taken as the reference. The testing data is associated to the class which has majority of the votes amongst the KNN [11]. When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance [12]. If the first instance is (a_1, a_2, a_3 … a_n) and the second instance is (b_1, b_2, b_3, … b_n), the distance between them is calculated by the following Equation (1):

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots (a_n - b_n)^2} \ldots (1)$$

KNN usually deals with continuous attributes however it can also deal with discrete attributes [13]. When dealing with discrete attributes if the attribute values for the two instances a_2, b_2 are different so the difference between them is equal to one otherwise it is equal to zero.

K-nearest Neighbor's algorithm

1. Determine the value of k.
2. Compute the distance between each record of the training set and the testing record.
3. Sort the neighbors in the increasing order of the distances.
4. Select the first k neighbors from the sorted list.
5. Check for the class, that majority of the neighbors belongs to and assign that class to the training data

B. SNMPTN

Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) is one way of acceptance of new students at Public University. SNMPTN is performed by each Public University using an integrated national system based on students' academic achievement in the form of report cards. The academic achievement that used are report cards semester 1 (one) up to semester 5 (five) for SMA / SMK / MA or equal with 3 (three) years study period [4]. Since 2011, each year only about 20% of the total applicants who received through SNMPTN from all over Indonesia? Although SNMPTN has the biggest quota among the selection of public university acceptances, but SNMPTN is not easy as well. The intense competition often makes students feel pessimistic to register SNMPTN [1].

C. Related Works Regarding to Prediction Applied K-nearest Neighbor Algorithm

There are some works related to prediction applied k-nearest neighbor algorithm. Alkhatib et al in 2013 [14] explored the stock prices to assist investors, management, decision makers, and users in making correct and informed investments decisions. This work utilized three attributes including closing price, low price, and high price to do the prediction. The results show that these three attributes (Closing price, Low price, and High price) can be determined as important parameter. The results were rational and reasonable. Depending on the actual

stock prices data, the prediction results were close to actual prices. Later, in 2016 Hasan et al [15] present an applied research on designing and developing a recommender system for graduate admission seekers which can help them to choose graduate school matching their entire academic profile. In this research, Hasan et al utilized CGPA, GRE, TOEFL_IELTS Score, Job experience, Research experience, Research area & Intended Outgoing country, Intended Semester, intended admission program as a training dataset.

The result shows 75% of accuracy and their proposed recommender system will recommend list of universities to applicants trying to pursue higher study abroad and eventually assist them to apply for graduate admission in appropriate universities. In 2016, Fitrianah et al [16] proposed a framework for identifying potential fishing zones (PFZs) based on a data-mining approach in the Eastern Indian Ocean. This work utilized a spatio-temporal clustering method to identify clusters of zones with data on the largest number of fish catch, which were then integrated with the sea surface temperature (SST) and the sea surface chlorophyll a (SSC) data derived from Moderate Resolution Imaging Spectro radiometer (MODIS) satellite imagery. The result gave an average accuracy of 87.11%, which showed that the proposed framework can be used effectively to determine PFZs. A prediction applied k-nearest algorithm has been done by Enriko et al in 2016 [17]. They utilized 8 Simplified Patient's Health of their dataset as parameters to predict heart disease. The parameters are Age, Sex, Chest pain, Resting blood pressure systolic, Resting blood pressure diastolic, Resting ECG, Resting heart rate, and Exercise induced angina. Experiments using 8 parameters with KNN shows 81.85% of accuracy. They can proof that 8 simple parameters are good enough to be used in heart attack prediction.

D. Related Works Regarding to SNMPTN

There is a work that related to SNMPTN. In 2016, Trisaputra et al analyzed the high school student who accepted through SNMPTN. This work utilized the average scores of semesters 3, 4, 5, 6 from the web that provides survey results with some Public Universities. The result shows that the K-nearest Neighbor method as a technique in data mining can be used for classifier in SNMPTN data. From several experiments, the best accuracy obtained is 83.3607%. But this work utilized data from public website which cannot be accounted for its accuracy.

III. METHODOLOGY

Overall the methodology is shown as in Figure 1 below. There are data collection, data preprocessing, implement of KNN and the last is development of application. The detailed step for implement KNN process is given in Figure 2 later.

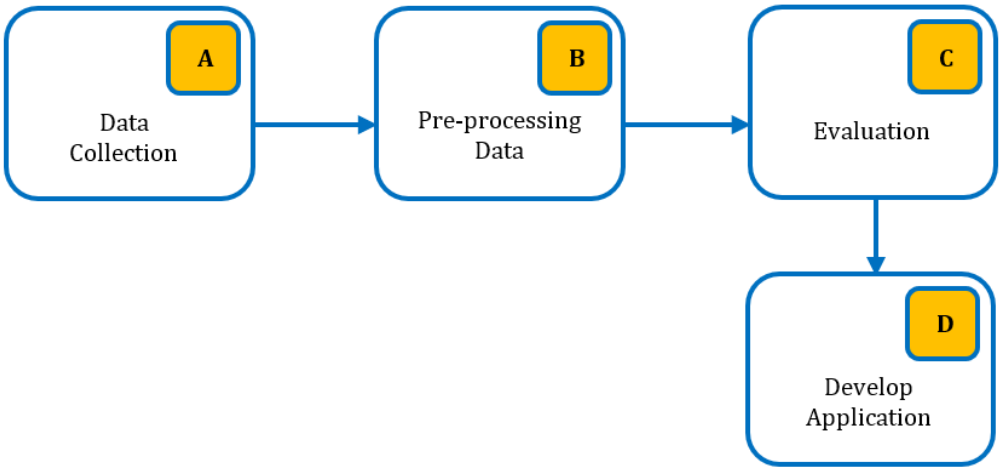


Fig. 1. Block Diagram of Research Method

A. Data Collection

The dataset that author has used in this work is taken from a public high school, SMAN 8 Jakarta alumni's data within 5 years (from 2013 - 2017). The weighted scores are calculated from the average scores from semesters 1 until 5. Alumni report card used as training data where there are two groups of alumni are accepted and they are not accepted by SNMPTN.

B. Data Preprocessing

Stages performed in the preprocesses include:

1. Selection of data and retrieval of data in accordance with the scope of the research. In this step, we select the alumni's data into two groups: science majoring alumni and social science majoring alumni. Then classifies them into the form of "Accepted" and "Not accepted" through SNMPTN that will be used as training data.
2. From all data of SMAN 8 Jakarta alumni's data, we only use some data that is only alumni who accepted and not accepted through SNMPTN at Universitas Indonesia and Institut Teknologi Bandung, which those data become the scope area of our research.
3. The subjects that will be used as training data according to the SNMPTN website [4] as follows:

- Science: Mathematics, Bahasa Indonesia, English, Chemistry, Physics, and Biology
- Social science: Mathematics, Bahasa Indonesia, English, Sociology, Economics, and Geography.

4. The resulting data is divided into 2 groups namely alumni of Science and Social science. The data consists of 200 records of alumni majoring in science and 100 records of alumni majoring in social science with each of them 10 attributes. Partial portion of alumni data majoring in Science and Social science as training data is given in Table 1 and 2 below.
5. Data cleaning, addressing missing data and data normalization. At this stage, we delete some data that does not have the completeness of the report cards with in semesters 1 until 5.
6. Transform the data, convert the data to a form or format appropriate to the device software used. We transformed the excel version of the document to the CSV version, to be used in applications WEKA

TABLE – I : PARTIAL PORTION OF SCIENCE MAJORING ALUMNI AS TRAINING DATA

Name	Major	Univ	Faculty	Avg 1	Avg 2	Avg 3	Avg 4	Avg 5	Status
Almira	Science	Universitas Indonesia	Fakultas Kedokteran	79	79	82	82	86	Accepted
Astrini	Science	Universitas Indonesia	Fakultas Teknik	83	83	82	85	85	Accepted
Devara	Science	Universitas Indonesia	Fakultas Teknik	79	78	80	80	83	Accepted
Irham	Science	Institut Teknologi Bandung	FITB	80	79	81	84	88	Accepted
Azaria	Science	Institut Teknologi Bandung	FITB	84	80	83	82	85	Accepted
Sydha	Science	-	-	85	85	81	83	84	Not Accepted
Fariz	Science	-	-	81	80	80	81	84	Not Accepted
Muhammad	Science	-	-	83	84	85	85	89	Not Accepted
…	…	…	…	…	…	…	…	…	…
Nur	Science	-	-	81	81	79	81	83	Not Accepted

TABLE- 2: PARTIAL PORTION OF SOCIAL SCIENCE MAJORING ALUMNI AS TRAINING DATA

Name	Major	Univ	Faculty	Avg 1	Avg 2	Avg 3	Avg 4	Avg 5	Status
Adinda	Social science	Universitas Indonesia	Fakultas Ekonomi & Bisnis	81	82	86	87	89	Accepted
Felino	Social science	Universitas Indonesia	Fakultas Hukum	78	81	83	85	87	Accepted
Gabriela	Social science	Universitas Indonesia	Fakultas Ilmu Sosial dan Imu, Politik	79	81	86	86	87	Accepted
Fairuza	Social science	Institut Teknologi Bandung	SBM	80	79	82	84	85	Accepted
Muhammad	Social science	Institut Teknologi Bandung	SBM	80	81	82	83	84	Accepted
Diadre	Social science	-	-	78	78	81	81	82	Not Accepted
Mohammad	Social science	-	-	75	77	80	81	83	Not Accepted
Prianza	Social science	-	-	72	76	77	79	79	Not Accepted
…	…	…	…	…	…	…	…	…	…
Student NON ITB21	Social science	-	-	82	81	84	86	88	Not Accepted

C. Evaluation

In order to evaluate the works, experiment utilizes the accuracy. Given the confusion matrix as on Table 3 below, the accuracy (Acc) value is obtained from Equation 2.

TABLE:3 - CONFUSSION MATIRX

		Predicted	
		Accepted	Not Accepted
Actual	Accepted	a	b
	Not Accepted	c	d

The accuracy (Acc) value is obtained from Equation 2,

$$\text{Accuracy} = \frac{(a+d)}{N} x100 \ldots (2)$$

Here, a is true positive, d is true negative, c is false positive and b is false negative. The value of N is the sum of a, b, c,and d [2].

D. Development of Application

According to Figure 1 showed in the previous section, we have designed and developed a web-based application which will predict student on SNMPTN acceptance. We use Relational Unified Process, or RUP is a framework for iterative software development process, controlled by the use case, based on the architecture by adding little by little (incremental) [18].

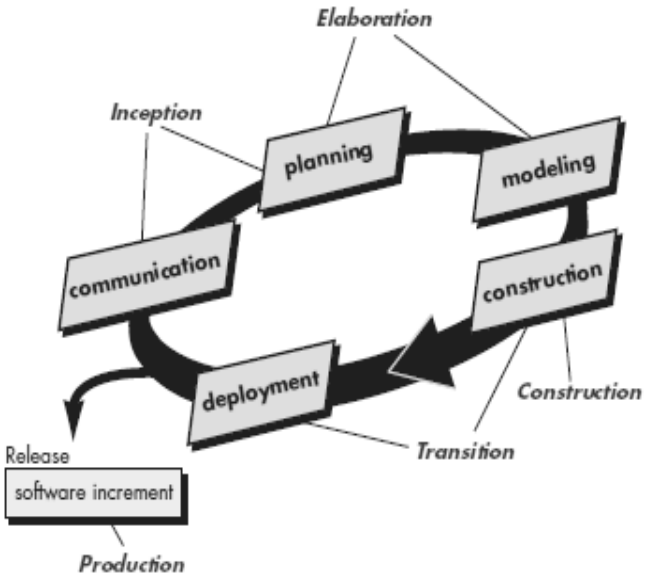


Fig. 2. The Unified Process[18]

The stages in RUP described in detail [5], as follows:

1) Stages of introduction (inception) from UP to discuss about communication with the users and discuss planning activities. At this stage, we determine the boundaries and scope of the object. Planning to determine the type of model that will be using in the software development process itself
2) Stages elaboration using communication activities and modeling owned generic process models. Stages of elaboration used to refine and develop the use case early developed in stages of inception. At this stage, we develop a project plan and analyze various requirements and risks.
3) Stages of construction in UP method is identical to the same activity, which is defined for the generic software process. Activities that occur in this phase include design, implementation, and of course testing software. In this phase we of course develop a project plan in the form of coding bit by bit (incremental).
4) Stages of transition to the UP using generic stages of construction activity and the first part of the submission and feedback components. Software is delivered to end users for beta testing and to get feedback from users on matters relating to defects program and the changes required. At this stage we make what has been modeled into a whole product. In this stage also performed such as, performance testing and create additional documentation.

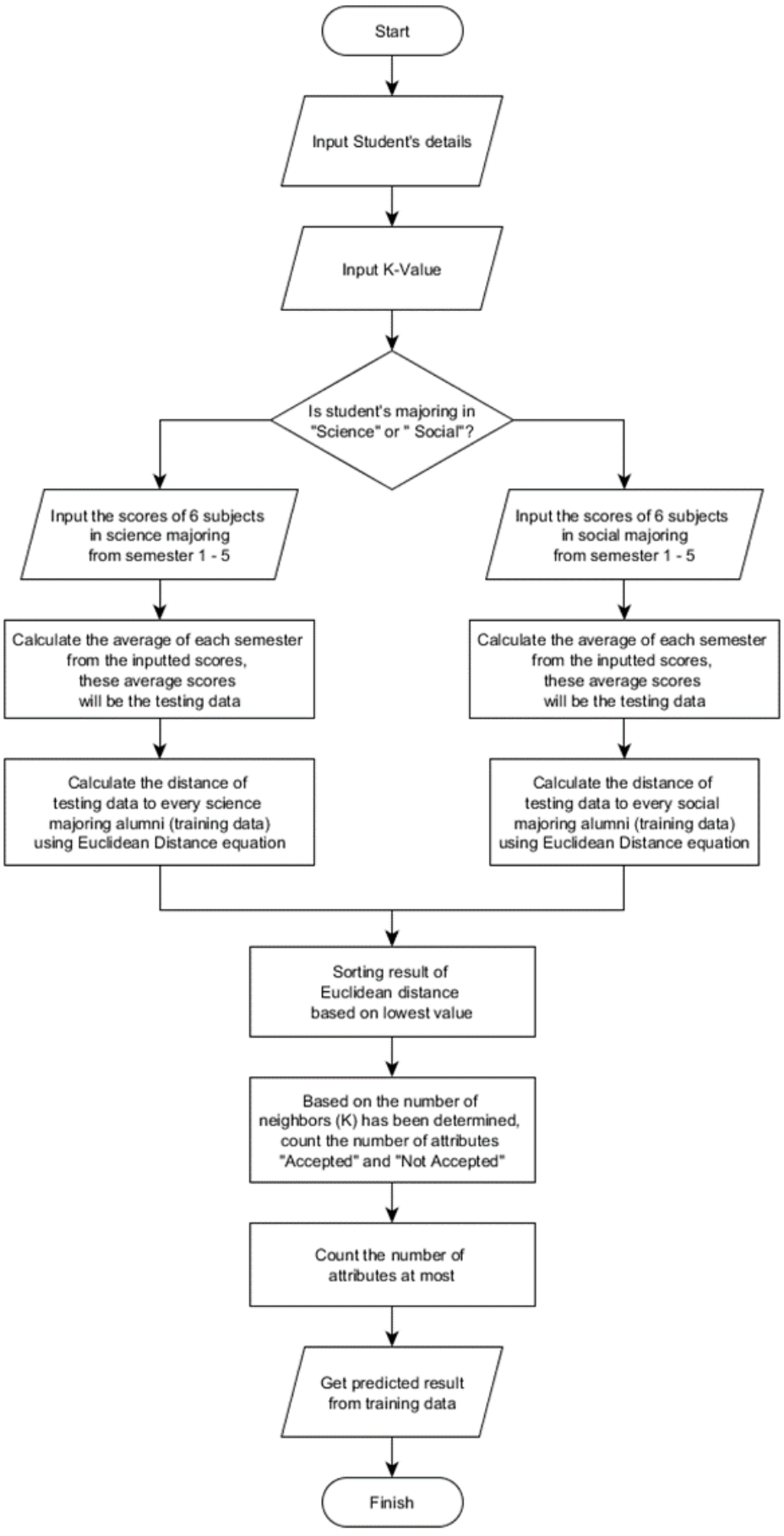


Fig. 3. Implementation of KNN to SNMPTN Prediction Application

Figure 3 shows the implementation of KNN to SNMPTN prediction application in this research. First, user input the testing data, which are the details of student and input k value that will be predicted. Second, user input all the scores of subjects according to their major within semester 1 until 5. If science majoring student, the subjects' scores inputted are Mathematics, Indonesian, English, Chemistry, Physics, and Biology, and if social science majoring student, the subjects' score inputted are Mathematics, Indonesian, English, Sociology, Economics, and Geography. Then, the application will calculate the average of each semester from all subjects' scores inputted according to student's major. These average scores of each semester (from semester 1 until 5) will be testing data. After that, the testing data inputted will be compared with training data according to student's major. If the testing data inputted is science majoring student, the testing data will be compared with science majoring alumni data, this also applies to the social science majoring. If all data are already inputted, the testing data will be calculated its Euclidean to all training data according to student's major. Then, sort the result of Euclidean distance based on lowest result. Based on the number of neighbors (K) which has been determined, count the number of attributes "Accepted" of "Not Accepted". Last of all, count the number of attributes at most and then we will get the predicted result.

IV. RESULT AND DISCUSSION

A. First Experiment: Using Science Majoring Alumni's Data as Training Data

Before we perform the KNN analysis, we may take a look at the parameters/variables. We want to know how important each variable is, in order to know which variables are more important than others. This information could be a reference when we want to do the parameter weighting in KNN. There are many ways to compute the variable importance, one of them is with Chi-Square attribute evaluation. With WEKA tool, we can determine the attribute importance with Chi-Square attribute evaluation with results informed in Table 4.

TABLE : 4 - RESULT OF 8 VARIABLES IMPORTANCE TEST OF SCIENCE MAJORING ALUMNI'S DATA USING WEKA

Rank	Score	Attribute
1	45.562	8 AVG5
2	32.516	6 AVG3
3	27.506	4 AVG1
4	25.257	7 AVG4
5	18.431	5 AVG2
6	0	3 Faculty
7	0	2 University
8	0	1 Major

From the test, now we know that 5 most important variables are: the average score of semester 1 (AVG1), the average score of semester 2 (AVG2), the average score of semester 3 (AVG3), the average score of semester 4 (AVG4), and the average score of semester 5 (AVG5), while other variables are concluded as less important. One more important thing in KNN algorithm is how to determine the optimum k parameter. First, k should be an odd number since we have to vote the nearest neighbors into two classes (Accepted or Not Accepted) so if we choose even numbers the result can be tied [19]. Second, many research have been done to determine optimum k parameter but there is no ultimate method to determine "optimum k parameters" so one method that can be used is to select k parameter using k=1 until k = square root of training data (k=1, k=3, k=7, ... k= $\sqrt{n}$) [20]. In this experiment, since the science majoring alumni's dataset consists of 200 records, and we also used 10-folds cross-validation as testing data, it means that 90% (or 180 records) are training data and 10% (or 20 records) are testing data. Thus, maximum number of k parameter for this experiment using science majoring alumni's data is square root of 180 which is 13.41. Finally, we can determine that the maximum number of k parameter is 13, for this experiment, as written in Table 5. In the first experiment, we used science majoring alumni's dataset of 200 records. We perform the analysis with WEKA tool. We compare the training data with two testing data models (training data itself and 10-folds cross validation) as testing data. In Table 5 shows the accuracy of each experiment using some K values.

TABLE : 5 - ACCURACY OF SCIENCE MAJORING ALUMNI'S DATA UTILIZED SOME K VALUES

K	Testing data mode	
	Evaluate on training data	10-fold cross-validation
03	80.0%	59.5%
05	77.0%	62.0%
07	72.5%	64.5%
09	71.5%	64.0%
11	73.5%	64.0%
13	72.0%	64.0%

From Table 5, we can see that the best result of the experiments using science majoring alumni's data as training data is 80% for evaluating the training data itself with k=3. Based on the results of experiments performed using WEKA shows that the value of k is very influential on the resulting accuracy as shown in Table 5. From this study, the best results were obtained from the smallest K value on the evaluation of the training data itself. This is because of the smaller the value of k, the less number of neighbors used for the new data classification process. Euclidean distance method is used to find the closeness between data, where the smaller the value, then the distance between the two of data is getting closer. Thus, when the value of k is small, then the neighbors who have the best data closeness are only used for the classification process.

B. Second Experiment: Using Social science Majoring Alumni's Data as Training Data

We do the same steps as science majoring alumni's data, now using social science majoring alumni's data. First, we conduct the Chi Squared attribute evaluation using WEKA tools to determine which variables are more important than others that mentioned in Table 6. From the test we found that top 5 variables are: the average score of semester 1 (AVG1), the average score of semester 2 (AVG2), the average score of semester 3 (AVG3), the average score of semester 4 (AVG4), and the average score of semester 5 (AVG5). We can state that these 5 variables are more important than 3 others. Then we do the KNN weighting experiments to check the accuracy. The results are concluded in Table 7.

TABLE : 6 - RESULT OF 8 VARIABLES IMPORTANCE TEST OF SOCIAL SCIENCEMAJORING ALUMNI'S DATA USING WEKA

Rank	Score	Attribute
1	56.69	4 AVG1
2	47.726	5 AVG2
3	41.319	6 AVG3
4	40.246	7 AVG4
5	39.472	8 AVG5
6	0	3 Faculty
7	0	2 University
8	0	1 Major

In this experiment, since the social science majoring alumni's dataset consists of 100 records, and we also used 10-folds cross-validation as testing data, it means that 90% (or 90 records) are training data and 10% (or 10 records) are testing data. Thus, maximum number of k parameter is square root of 90 which is 9.48. Finally, we can determine that the maximum number of k parameter for this experiment using social science majoring alumni's is 9, as written in Table 7. In the second experiment, we used social science majoring alumni's dataset of 100 records. We perform the analysis with WEKA tool. We compare the training data with two testing data models (training data itself and 10-folds cross validation) as testing data. In Table 7 shows the accuracy of each experiment using some K values.

TABLE : 7 - ACCURACY OF SOCIAL SCIENCE MAJORING ALUMNI'S DATA UTILIZED SOME K VALUES

K	Testing data mode	
	Evaluate on training data	10-fold cross-validation
3	89%	77%
5	82%	71%
7	80%	72%
9	77%	76%

From Table 7, we can see that the best result of the experiments using social science majoring alumni's data as training data is 89% for evaluating the training data itself with k=3. Based on the results of experiments performed using WEKA shows that the value of k is very influential on the resulting accuracy as shown in Table 7. From this study, the best results were obtained from the smallest K value on the evaluation of the training data itself. This is because of the smaller the value of k, the less number of neighbors used for the new data classification process.

C. Result of Implementation of KNN to SNMPTN Prediction Application

According to flow diagram in the figure 1, we have designed and developed a web based application utilized AdminLTE as web template and CodeIgniter as PHP framework which open source and uses MVC (Model, View, Controller) methods. This SNMPTN prediction application can be a tool help the school in predicting their students in public universities through SNMPTN.

1) Login Screen

The login screen of SNMPTN Prediction application (in Figure 4) includes two boxes to fill username and password to enter and access into the main page.

Fig. 4. Login Screen

2) Prediction Screen/Home Page

As soon as the user clicks on "Prediction Page" button, a screen will appear where user requires to fill all his/her necessary details to predict him/her suitable public university matching his/her given details in this UI. In this page, the user chooses the major of student, whether the student's major is science or social science. The prediction process will adjust to the student's major, if student to be predicted is majoring in science, then the training data will be used for predicting is science majoring alumni's data, as well as social science major. The Prediction screen of SNMPTN also includes K-Value input which will be predicted

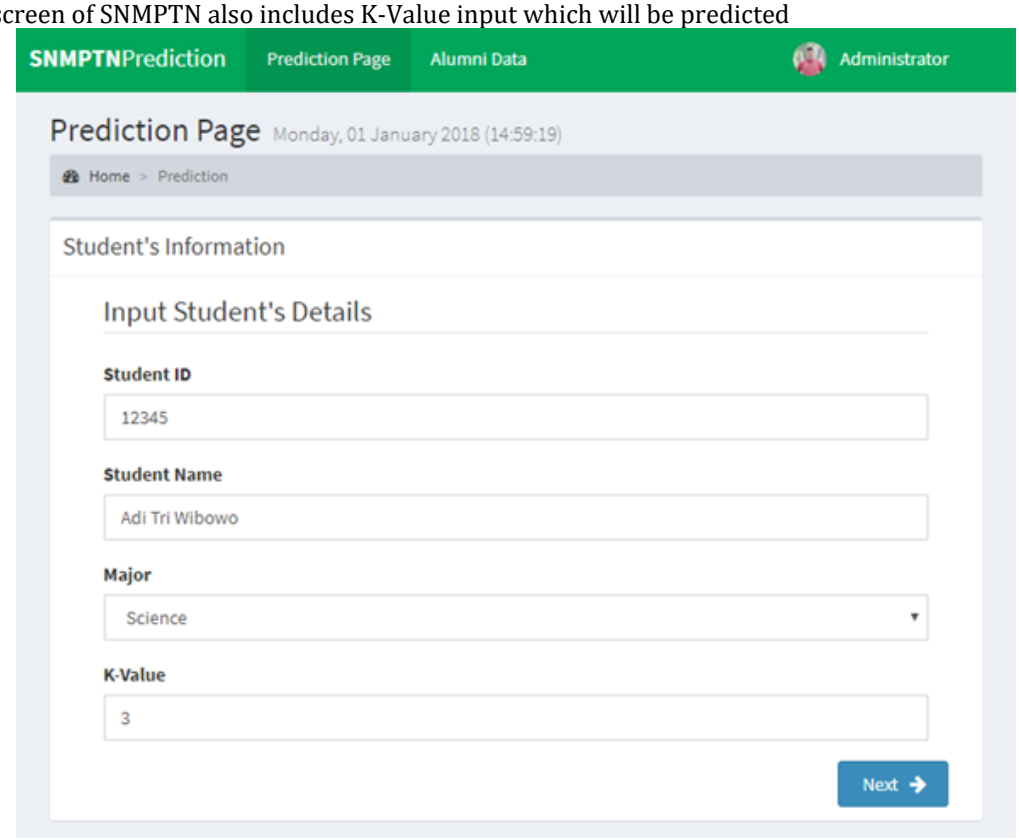


Fig. 5. Home Screen

3) Student's Score Input Form

The student's score input form of SNMPTN includes student's scores from semester 1 until semester 5 which will be predicted. If the student to be predicted is majoring in science, then the subjects that will appear are Mathematics, Bahasa Indonesia, English, Chemistry, Physics, and Biology. If the student to be predicted is majoring in social science, then the subjects that will appear are Mathematics, Bahasa Indonesia, English, Sociology, Economics, and Geography. In this process, the application will count the average of each semester. These scores will become the testing data that is going to be predicted.

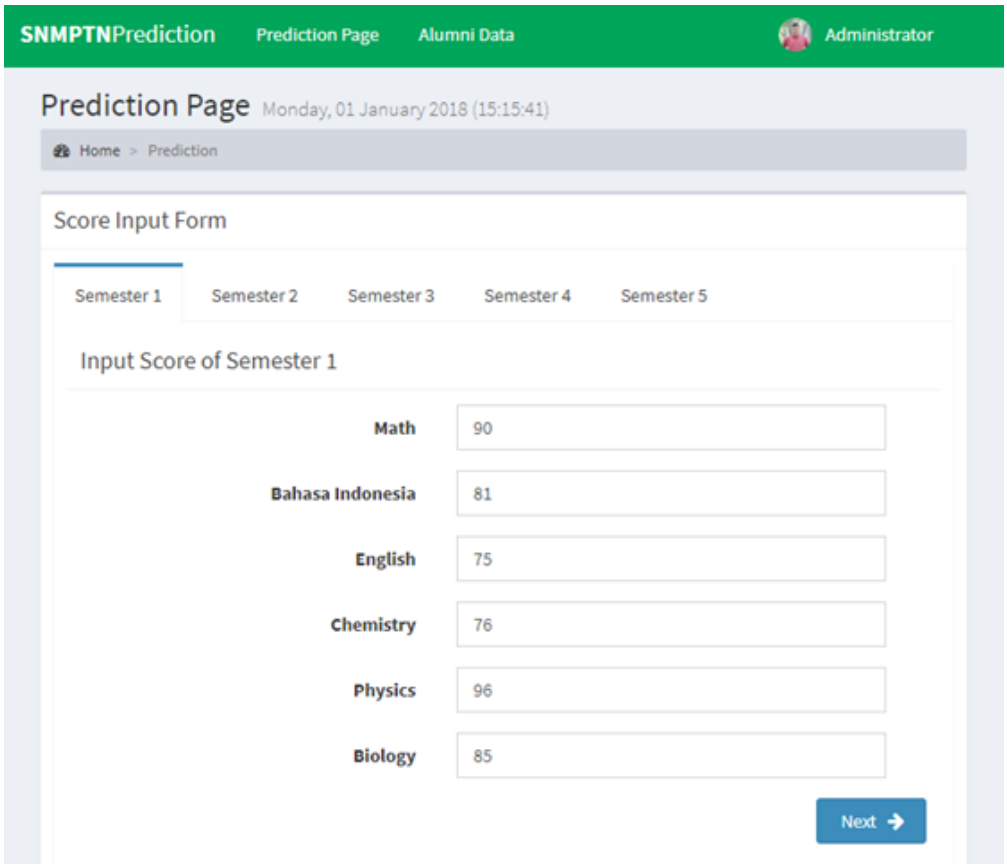
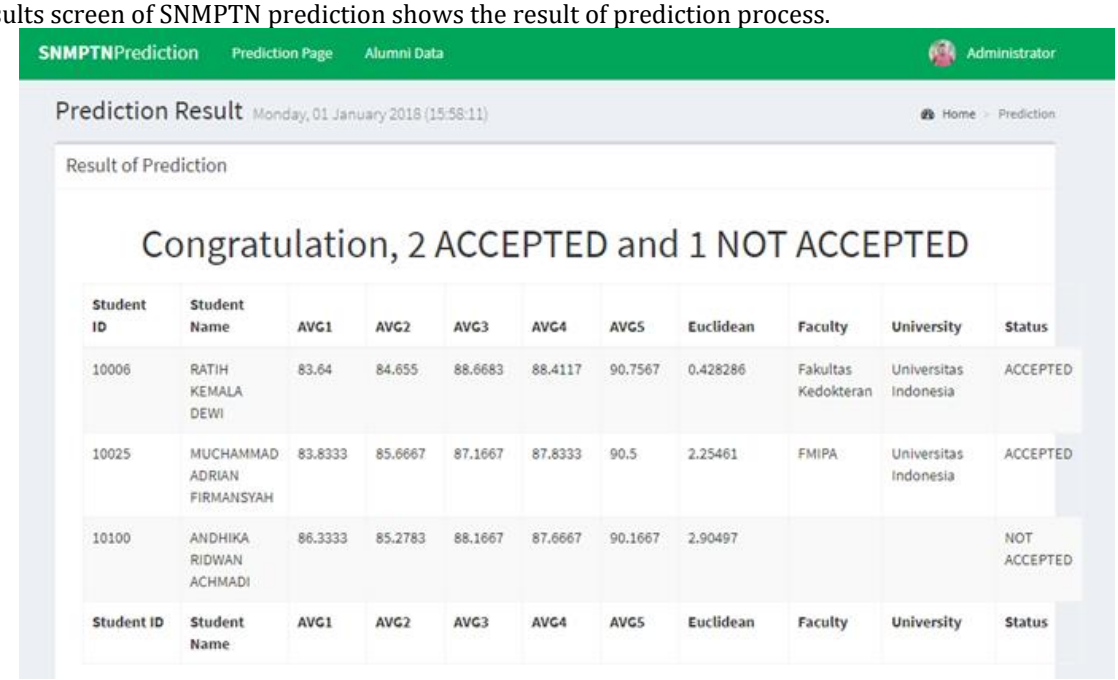


Fig. 6. Score Input form

4) Result Screen

The results screen of SNMPTN prediction shows the result of prediction process.



Student ID	Student Name	AVG1	AVG2	AVG3	AVG4	AVG5	Euclidean	Faculty	University	Status
10006	RATIH KEMALA DEWI	83.64	84.655	88.6683	88.4117	90.7567	0.428286	Fakultas Kedokteran	Universitas Indonesia	ACCEPTED
10025	MUCHAMMAD ADRIAN FIRMANSYAH	83.8333	85.6667	87.1667	87.8333	90.5	2.25461	FMIPA	Universitas Indonesia	ACCEPTED
10100	ANDHIKA RIDWAN ACHMADI	86.3333	85.2783	88.1667	87.6667	90.1667	2.90497			NOT ACCEPTED
Student ID	Student Name	AVG1	AVG2	AVG3	AVG4	AVG5	Euclidean	Faculty	University	Status

Fig. 7. Result Screen

5) Alumni Data/Training Data Screen

In the alumni data or training data screen, user enables to input the score of alumni who accepted and not accepted through SNMPTN as training data.

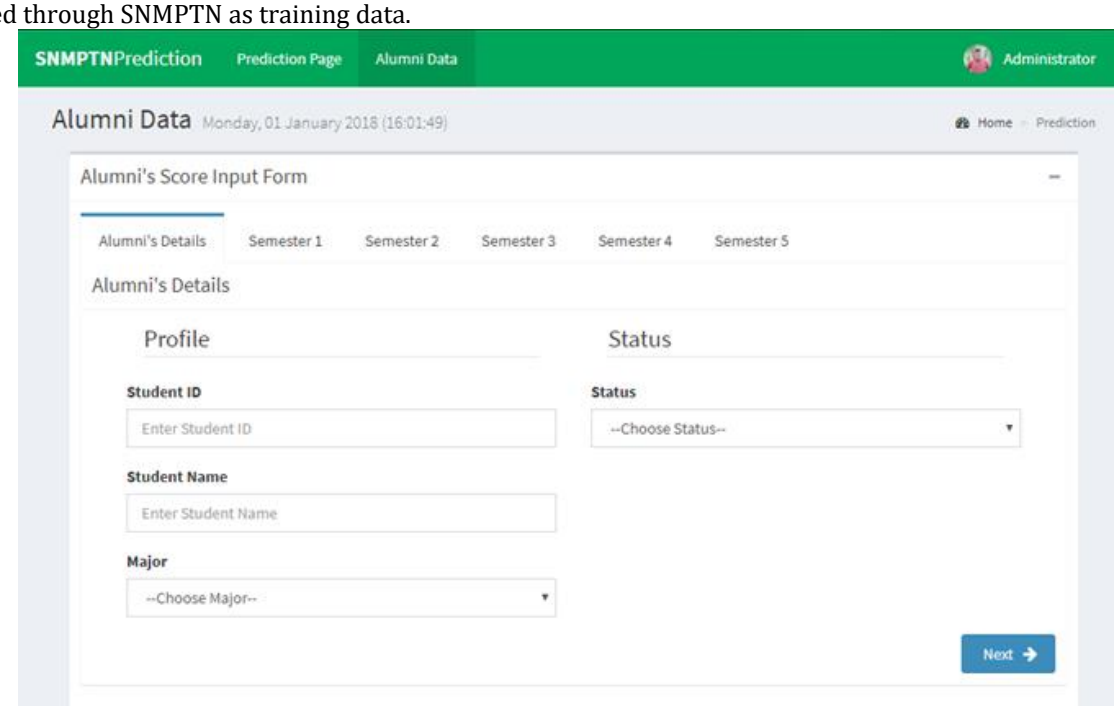


Fig. 8. Training Data

V. IMPLICATION

The classification results show a prominent accuracy. It is functioned as the classification method in the web application developed. Some parameters are also identified in predicting good results. These parameters are the average score from the semester 1 until semester 5 can be included in the predicting as the learning data. The SNMPTN web application can be utilized an alternative tool that assist high schools in predicting the student acceptance to public universities through SNMPTN. Yet, it also can assist students in determining their choice and provide recommendationlist of universities that can be enrolled. The implication of the study is, it will be easier for schools to direct and manage students in order to focus to the targeted universities.

VI. CONCLUSION

Prediction classification with K-Nearest neighbor algorithm successfully done with good results. From this study we get the 5 optimal parameters out of 8 parameters to gain good prediction, they are the average score of semester 1, the average score of semester 2, the average score of semester 3, the average score of semester 4, and the average score of semester 5. The best parameter to predict the Science majoring alumni data is the average score from semester 5 with score 45.56 and the best parameter to predict the Social Science majoring alumni data is the average score from semester 1 with science 56.69. The study shows good results, as shown in accuracy obtained are 80% for evaluating the Science majoring alumni's data itself with K=3 and 89% for evaluating the Social Science majoring alumni's data itself with K=3. Thus, the prediction application developed is sufficient and suitable for predicting SNMPTN. Last of all, this SNMPTN prediction application can be a tool help the schools in predicting their students in public universities through SNMPTN.

REFERENCES

1. Y. Trisaputra, "Klasifikasi Profil Siswa SMA / SMK yang Masuk PTN (Perguruan Tinggi Negeri) dengan k-Nearest Neighbor," no. August, pp. 0–15, 2016.
2. D. Fitrianah, N. H. Praptono, A. N. Hidayanto, and A. M. Arymurthy, "Feature Exploration for Prediction of Potential Tuna Fishing Zones," Int. J. Inf. Electron. Eng., vol. 5, no. 4, pp. 270–274, 2015.
3. D. Sonawane, "Prediction Using Back Propagation and k- Nearest Neighbor (k-NN) Algorithm," Int. J. Innov. Reserach Comput. Commun. Eng., vol. 3, no. 4, pp. 3209–3213, 2015.
4. "Informasi Umum SNMPTN 2017." [Online]. Available: http://snmptn.ac.id/informasi.html?1426322267. [Accessed: 25-Dec-2017].
5. H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k-Nearest Neighbour Classification in Medical Data Mining," Int. J. Inf. Commun. Technol. Res., vol. 4, no. 4, pp. 121–128, 2014.
6. M. Shouman, T. Turner, and R. Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," Int. J. Inf. Educ. Technol., vol. 2, no. 3, pp. 220–223, 2012.

7. F. Barigou, "Improving k-nearest neighbor efficiency for text categorization," Neural Netw. World, vol. 26, no. 1, pp. 45–65, 2016.
8. H. R. Shahraki, S. Pourahmad, and N. Zare, "? Important Neighbors : A Novel Approach to Binary Classification in High Dimensional Data," vol. 2017, 2017.
9. N. Suguna and K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm," Int. J. Comput. Sci. Issues, vol. 7, no. 4, pp. 18–21, 2013.
10. D. J. Bora and A. K. Gupta, "Effect of Different Distance Measures on the Performance of K-Means Algorithm : An Experimental Study in Matlab," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 2501–2506, 2014.
11. A. Giri, M. V. V. Bhagavath, B. Pruthvi, and N. Dubey, "A Placement Prediction System using k-nearest neighbors classifier," Proc. - 2016 2nd Int. Conf. Cogn. Comput. Inf. Process. CCIP 2016, pp. 3–6, 2016.
12. K. Saxena, Z. Khan, and S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," Int. J. Comput. Sci. Trends Technol., vol. 2, no. 4, pp. 36–43, 2014.
13. V. K. D.Shanmugasundaram, "The Comparative Study for Diagnosing Heart Disease Using KNN and Naïve Bayes," pp. 9–19, 2015.
14. K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi, "Stock Price Prediction Using K-Nearest Neighbor Algorithm," Int. J. Business, Humanit. Technol., vol. 3, no. 3, pp. 32–44, 2013.
15. M. Hasan, S. Ahmed, D. M. Abdullah, and M. S. Rahman, "Graduate school recommender system: Assisting admission seekers to apply for graduate studies in appropriate graduate schools," 2016 5th Int. Conf. Informatics, Electron. Vision, ICIEV 2016, pp. 502–507, 2016.
16. D. Fitrianah, A. N. Hidayanto, J. L. Gaol, H. Fahmi, and A. M. Arymurthy, "A Spatio-Temporal Data-Mining Approach for Identification of Potential Fishing Zones Based on Oceanographic Characteristics in the Eastern Indian Ocean," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 9, no. 8, pp. 3720–3728, 2016.
17. I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient â€™ s Health Parameters," vol. 8, no. 12, 2016.
18. R. S. Pressman, Software Engineering A Practitioner's Approach 7th Ed - Roger S. Pressman. 2009.
19. M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid-Ahmed, "Investigating the performance of Naive- Bayes classifiers and K- nearest neighbor classifiers," 2007 Int. Conf. Converg. Inf. Technol. ICCIT 2007, no. April, pp. 1541–1546, 2013.
20. A. B. Hassanat, M. A. Abbadi, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," Int. J. Comput. Sci. Inf. Secur., vol. 12, no. 8, pp. 33–39, 2014.

