Comparison of K-Nearest Neighbour and support vector machine for choosing senior high school

M Irfan1**, A R Nurhidayat4, A Wahana1, D S Maylawati1,2 and M A Ramdhani1

1Department of Informatics, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia
2Department of Informatics, Sekolah High Tekologi Garut, Garut, Indonesia

Corresponding author’s email: irfan.bahaf@uinsgd.ac.id

Abstract. The aim of this research is to compare K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) algorithm which are used for choosing senior high school recommendation. As we know that education is an important aspect in the development of a nation. The methodology that used in this research is data mining through Knowledge Discovery in Database (KDD) stages, which consist of cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation. KNN and SVM are the most common algorithms used in data mining and decision support system. Either KNN or SVM in this research used for classifying the type of senior high school by input parameters, among others national examination score, student interest, and counsellor suggestion. Based on the experiment with several training and testing data, the result shows that SVM is better than KNN. SVM has an accuracy value around 97.1%, while KNN has an accuracy value around 88.5%. And also, SVM has processing time faster than KNN.

1. Introduction
Education is the most important sector in the development of a nation. A nation is said to be an advanced nation if the education level of its citizens is good. The aim and function of national education is to develop the ability and build the dignified national civilization in order to educate the life of the nation, to develop the potential of students to become believers and fear of God Almighty, noble, healthy, knowledgeable, capable, creative, independent and become a democratic and responsible citizen.

To realize the aim of national education optimally, each student needs to take formal education at least until students take senior high school and it is better to continue to university. The level of formal education starts with basic education (elementary school), then continues with secondary education (junior high school and senior high school). Senior high school in Indonesia consists of general secondary education and vocational secondary education. General secondary education such as senior high school (SMA), Madrasah Aliyah (Islamic senior high school/ MA), vocational high school (SMK), vocational Madrasah Aliyah (MAK), or other equivalent secondary education.

In line with the explanation above, after graduating from junior high school, every grade IX student (third-year student) should continue their education to senior high school, Madrasah Aliyah, vocational high school, or another equivalent school. Commonly, class IX students who take junior high school education will certainly faced with the problem of choosing senior high schools, both general senior high school, and vocational high schools. Choosing a right senior high school is an important factor because it relates to the future of students. In determining of senior high school, a
variety of considerations needed, one of them is the student concerned. Because every student has the right to choose a school according to his talent. Therefore, it is necessary to design an expert system that can help students to determine their advanced studies by their potentials. Expert system is a system that stores expert knowledge, not used to replace the role of an expert, but to help solve problems like an expert [1-3].

Similar studies that had been conducted, including: bank credit risk analysis with k-nearest-neighbour classifier: case of Tunisian banks[4]; the implementation of Nearest Neighbour and CBR algorithms on expert systems of sexual behavior deviations [5]; ace recognition uses SVM Multi Kernel with increased learning [6]; the comparation of text mining with Naïve Bayes classifier, nearest neighbour, and decision tree to detect Indonesian swear words on Twitter [7]; classification system of philodendron leaf ornamental plants using the K-Nearest Neighbour (KNN) method based on the Hue Saturation Value (HSC) [8]; and the comparison of Naïve Bayes Classifier with Nearest Neighbour for identification of eye diseases [9]. This study uses as well as comparing the use of the KNN algorithm with SVM for the determination of senior high school for junior high school students

2. Methods
This study uses a prototype model as software development life cycle. Prototyping is a fast development and testing of the model (prototype) of a software through an interactive process [10-12]. Then, for modelling the analysis and design of this study use the Unified Modelling Language (UML) approach that parses the system with visual models [13]. The algorithm that chosen as the model implemented in the system is a data mining mechanism. Data Mining is a useful technology to find the most important information (insight of data) [6]. The data mining methodology used in this study uses the Knowledge Discovery in Database (KDD) model. Beginning from database collection that used for cleaning and integration data, preparing flat files in data warehouse then conducting selection and transformation from data warehouse, conducting data mining itself, and the last find the knowledge from evaluation and presentation as the result of data mining process.

In detail, the phases of KDD, among others [14]: Data Cleaning: is a process to eliminate data that contains noise, inconsistent data, or redundant data; Data Integration: in this stage, the data will be merged from various sources; Data Selection: is a process to select a relevant data from database; Data Transformation: in this stage, the data will be transformed into the suitable format for the needs of data mining process; Data Mining: is an important process where the pattern of data will be extracted using various methods or algorithms; Pattern Evaluation: the patterns that have been produced from data mining process will be evaluated based on certain metrics for finding interesting pattern; and Knowledge Presentation: this is the last stage where the knowledge that have found will be visualized to described the important information more clearly and easily. System reliability in this research is reached by analytical, logical, conceptual, and operational verification [15].

3. Result and Discussion

3.1. Problem Analysis
This research is an information system design to help junior high school students to choose a senior high school based on their potentials. Information system combine the technology and human activities [16], such as used to support decision making [17]. Information systems are systems that manage the organized data [1], that built based on user requirements [18], information systems have a good flexibility that allows them to be developed into better systems [19], and capable of solving complex data [20]. The expert system can also be said as an intelligent information system that has expert knowledge and solve the problem as like as an expert [21]. Based on much previous research, information system has advantages such as good data accessibility [22], efficient processing time [23], accurate [24], precise support decision [25], more economical [26], widely used [27], increase the understanding of user [2], improve productivity [28], present data and information well [29], and as media to save data [30].

In this study, data mining is used as a technique to recommend selection of senior high school, and counsellors are needed as an expert to give their knowledge and judgments for the functioning quality
of the system in solving the problem. Data Mining is defined as a series of processes to get knowledge or patterns from a data set [31]. The development of data mining cannot be separated from the advancement of information technology that allows large amounts of data to accumulate. As data mining is increasingly needed, several classification algorithms appear to process large amounts of data [32].

Classification is a supervised learning process. Classification is defined as a form of data analysis to extract a model that will be used to predict class labels [14]. Classification methods commonly used such as Decision Tree, KNN, Naïve Bayes, Artificial Neural Network, and SVM. KNN is a classification method that determines the label (class) of a new object based on the class that is a majority from k-neighbour in the training set. KNN is a classification algorithm based on analogy, which compares testing data with training data that are close to and have similarities to the test data. [33]. While SVM Algorithm is a classification method in machine learning with Structural Risk Minimization (SRM) principle that aim to find the best hyperplane that separates two classes in the input space [34]. Either KNN or SVM algorithms has advantages and disadvantages. Therefore, in this study will make a comparison between KNN and SVM algorithms to obtain which algorithm that has better accuracy in the classification that determines the senior high schools’ recommendation for class IX junior high school students (the case study of this research at Madrasah Tsanawiyah/ MTS/ Islamic Junior High School Nurul Anwar). So, the comparison parameters of the KNN and SVM algorithms are the accuracy of the classification result and the processing time.

3.2. Data Mining

3.2.1. K-Nearest Neighbour Algorithm. The KNN algorithm is a method for classifying objects based on learning data that are closest to the object. Learning data is projected to be a multi-dimensional space, where each dimension represents the features of the data. This space divided into sections based on the classification of learning data. A point in this space is marked with class c, if class c is the classification that most often found in the k-neighbour of the closest to that point [35].

KNN is a supervised method, where the results of querying new instances classified according to the majority of categories on the KNN. In the training phase, this algorithm only stores feature vectors and classifications of training sample data. In the classification phase, the same features calculated for data testing (the classification is unknown). The distance from this new vector to all sample training vectors calculated, and the closest k will taken. The new point of classification that is already predicted to included in the most classification of these points. Equation (1) is used to calculate the similarity of the data in k-neighbour.

\[
\text{Similarity} (T, S) = \frac{\sum_{i=1}^{n} \text{Sim}(K(T), K(S))w_i}{\sum_{i=1}^{n} w}
\]

Where:  
T = testing data  
S = training data  
n = total of criteria  
w = weight of criteria  
Sim (K(T), K(S)) = Similarity value or criteria distance between case target and source target

3.2.2. Support Vector Machine Algorithm. SVM is a learning machine method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the optimal hyperplane that separates two classes in the input space [6]. The optimal hyperplane is a hyperplane that located in the middle between two sets of objects from two classes. The optimal hyperplane as a separator between the two classes can be found by measuring the margin of the hyperplane and finding its maximum point. The margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called support vector [34]. Figure 1 describes the concept of SVM to find the optimal hyperplane.
3.3. Implementation and Experiment

3.3.1. K-Nearest Neighbour Algorithm. The processes of KNN Algorithm are among others: Determine k parameter (total closest neighbour); Calculate the square of the Euclidean distance of the object against the given training data; Sort the result of (b) from the highest value into a lowest value; Gather Y category (nearest neighbour classification based on k value); and by using the closest majority category, the category of the class of objects can predicted. Table 1 until Table 4 presents the calculation of closeness value between an object with 10 data sample. Where SMA is general senior high school, SMK is vocational high school, MA is Madrasah Aliyah. The percentage of each parameter given by school academics and counselor.

Table 1. Value closeness with the KNN algorithm

| National Exam Score (30%) | Student Interest (50%) | Counsellor Suggestion (20%) |
|---------------------------|------------------------|-----------------------------|
| Training Data | Testing Data | Closeness Value | Training Data | Testing Data | Closeness Value | Training Data | Testing Data | Closeness Value |
| High | High | 1 | SMK | SMK | 1 | SMK | SMK | 1 |
| High | Middle | 0.5 | SMK | SMA | 0.5 | SMK | SMA | 0.5 |
| High | Low | 0 | SMK | MA | 0 | SMK | MA | 0 |
| Middle | High | 0.5 | SMA | SMK | 0.5 | SMA | SMK | 0.5 |
| Middle | Middle | 1 | SMA | SMA | 1 | SMA | SMA | 1 |
| Middle | Low | 0.5 | SMA | MA | 0.5 | SMA | MA | 0.5 |
| Low | High | 0 | MA | SMK | 0 | MA | SMK | 0 |
| Low | Middle | 0.5 | MA | SMA | 0.5 | MA | SMA | 0.5 |
| Low | Low | 1 | MA | MA | 1 | MA | MA | 1 |

Table 2. Training data sample for KNN Algorithm

| No | Name | National Exam Score | Student Interest | Counsellor Suggestion | Goal |
|----|------|---------------------|------------------|-----------------------|------|
| 1  | Ade Y | Middle | SMK | SMK | SMK |
| 2  | Anggi | Middle | SMK | SMK | SMK |
| 3  | Laela | High | SMA | SMA | SMA |
| 4  | Indra | Middle | SMK | SMK | SMK |
| 5  | Wati | Middle | SMA | SMA | SMA |
| 6  | Risma | Middle | SMA | SMA | SMA |
| 7  | Siti P | Middle | SMA | SMA | SMA |
| 8  | Faysal | Low | SMA | MA | MA |
| 9  | Lastri | Middle | SMA | SMA | SMA |
| 10 | Hana | Middle | SMK | SMK | SMK |
Furthermore, senior high school recommendations in 3.3.2.

Table 3. Calculation of closeness value eith KNN Algorithm

| Closeness | National Exam Score | Student Interest | Counsellor Suggestion | Closeness Value | Class   |
|-----------|---------------------|------------------|-----------------------|----------------|--------|
| data1     | S R 0.5 0.3 0.2 SMK | SMK 1 0.5 0.5 SMK | MA 0 0.2 0 0.65    |                |        |
| data2     | S R 0.5 0.3 0.2 SMK | SMK 1 0.5 0.5 SMK | MA 0 0.2 0 0.65    |                |        |
| data3     | T R 0 0.3 0 SMA    | SMK 0.5 0.3 SMK | MA 0 0.2 0 0.25    |                |        |
| data4     | S R 0.5 0.3 0.2 SMK | SMK 1 0.5 0.5 SMK | MA 0 0.2 0 0.65    |                |        |
| data5     | S R 0.5 0.3 0.2 SMA | SMA 0.5 0.3 SMA | MA 0.5 0.2 0.1 0.50|                |        |
| data6     | S R 0.5 0.3 0.2 SMA | SMA 0.5 0.3 SMA | MA 0.5 0.3 0.3     |                |        |
| data7     | S R 0.5 0.3 0.2 SMK | SMK 1 0.5 0.5 SMK | MA 0 0.2 0 0.65    |                |        |
| data8     | R R 1 0.3 0.3 SMA   | SMK 0.5 0.3 SMK | MA 1 0.2 0.2 0.75  |                |        |
| data9     | S R 0.5 0.3 0.2 SMA | SMA 0.5 0.3 SMA | MA 0.5 0.2 0.1 0.50|                |        |
| data10    | S R 0.5 0.3 0.2 SMK | SMK 1 0.5 0.5 SMK | MA 0 0.2 0 0.65    |                |        |

Legend: \( R = \) Low; \( S = \) Middle; \( T = \) High; \( N1 = 1st \) test value \( N2 = 2nd \) test value; \( K = \) Closeness; \( BN = \) Weight value \( HN = K \ast BN \)

Table 4. The Result of KNN Algorithm

| No | Name | National Exam Score | Student Interest | Counsellor Suggestion | Closeness Value | Class   |
|----|------|---------------------|------------------|-----------------------|----------------|--------|
| 1  | Faysal | Low            | SMA              | MA                    | 0.75           | MA     |
| 2  | Ade   | Middle            | SMK              | SMK                   | 0.65           | SMK    |
| 3  | Anggi | Middle            | SMK              | SMK                   | 0.65           | SMK    |
| 4  | Indra | Middle            | SMK              | SMK                   | 0.65           | SMK    |
| 5  | Siti  | Middle            | SMK              | SMK                   | 0.65           | SMK    |
| 6  | Hana  | Middle            | SMK              | SMK                   | 0.65           | SMK    |
| 7  | Lastri| Middle           | SMA              | SMA                   | 0.50           | SMA    |
| 8  | Wati  | Middle            | SMA              | SMA                   | 0.50           | SMA    |
| 9  | Risma | Middle            | SMA              | SMA                   | 0.40           | SMA    |
| 10 | Laela | High              | SMA              | SMA                   | 0.25           | SMA    |

3.3.2. Support Vector Machine. Implementation of SVM algorithm is the same as the sample that used in KNN algorithm. The process and the result of SVM presented in Table 5 until Table 7. Furthermore, senior high school recommendations class is separated with three dividing lines, among others \( Y_1, Y_2, \) and \( Y_3 \) (Figure 2)

Table 5. Data Training of SVM algorithm (the weight value can see in Table 6)

| No | Name  | Weight of National Exam Score | Weight of Student Interest | Weight of Counsellor Suggestion |
|----|-------|-------------------------------|---------------------------|-------------------------------|
| 1  | Ade   | 0.5                           | 1                         | 1                             |
| 2  | Anggi | 0.5                           | 1                         | 1                             |
| 3  | Laela | 1                             | 0.5                       | 1                             |
| 4  | Indra | 0.5                           | 1                         | 1                             |
| 5  | Wati  | 0.5                           | 0.5                       | 0.5                           |
| 6  | Risma | 0.5                           | 0.5                       | 1                             |
| 7  | Siti  | 0.5                           | 1                         | 1                             |
| 8  | Faysal| 0                             | 0.5                       | 0                             |
| 9  | Lastri| 0.5                           | 0.5                       | 0.5                           |
| 10 | Hana  | 0.5                           | 1                         | 1                             |

Table 6. Criteria for SVM algorithm

| Criteria   | Letter Value | Weight Value |
|------------|--------------|--------------|
| National Exam Score | High | 1        |
| Low        | 0            |
| Student Interest | SMK | 1        |
| SMA        | 0.5          |
| MA         | 0            |
| Counsellor | SMK | 1        |
| Suggestion | SMA | 0.5     |
| MA         | 0            |
3.3.3. Result of Accuracy and Processing Time. The experiment of accuracy and processing time of KNN and SVM algorithm use were conducted using 35 student’s data. The results of accuracy and processing time using KNN classifier with 30 data in the correct class and 4 data in the incorrect class. Based on the result, the KNN has an accuracy of around 85.71%. While, for the SVM algorithms classify 34 in the correct class and 1 data in the incorrect class, so that SVM has an accuracy around 97.10%. And also, the processing time for SVM is faster than KNN, SVM has 0.001383 seconds while KNN has 0.003690 seconds of processing time.

4. Conclusion

Based on the experiment of this research, the implementation SVM, and KNN algorithm in classify the recommendation of the senior high school selection, the SVM algorithm has better accuracy and processing time compared to the KNN algorithm, which in this study uses data samples from class IX students of MTs Nurul Anwar, Sindangkasih, Ciamis. This study recommends the use of the SMV algorithm (compared to the KNN algorithm) for expert systems in determining senior high schools or for another similar study.

5. References

[1] Maylawati D S, Darmalaksana W and Ramdhani M A 2018 Systematic Design of Expert System Using Unified Modelling Language. In IOP Conference Series: Materials Science and Engineering 288 1 012047

[2] Maylawati D S, Ramdhani M A, Zulfikar W B, Taufik, I and Darmalaksana W 2017 Expert system for predicting the early pregnancy with disorders using artificial neural network

[3] Khosyi’Ah, S, Irfan M, Maylawati, D S and Mukhlas O S 2018 Analysis of Rules for Islamic Inheritance Law in Indonesia Using Hybrid Rule Based Learning. In IOP Conference Series: Materials Science and Engineering 288 1 012133

[4] Abdelmoula A K 2015 Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. Accounting and Management Information Systems 14 1 79
[5] Salamun S 2018 Penerapan Algoritma Nearest Neighbor dan CBR pada Expert System Penyimpangan Perilaku Sexual. *Jurnal Online Informatika* **2** 63-70

[6] Athoiilah M 2018 Pengenalan Wajah Menggunakan SVM Multi Kernel dengan Pembelajaran yang Bertambah. *Jurnal Online Informatika* **2** 84-91

[7] Zulfikar W B, Irfan M, Alam C N and Indra M 2017 The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter. In *Cyber and IT Service Management (CITSM)*, 2017 5th International Conference on 1-5

[8] Syahid D, Jumadi, J and Nursantika D 2016 Sistem Klasifikasi Jenis Tanaman Hias Daun Philodendron Menggunakan Metode K-Nearest Neighbor (KNN) Berdasarkan Nilai Hue, Saturation, Value (HSV). *Jurnal Online Informatika*, **1** 20-23

[9] Zulfikar W B and Lukman N 2016 Perbandingan Naive Bayes Classifier dengan Nearest Neighbor untuk Identifikasi Penyakit Mata. *Jurnal Online Informatika*, **1** 82-86

[10] Rosa M 2013 A Shalahuddin, Rekayasa Perangkat Lunak. *Bandung: Modula.*

[11] Dewanto IJ 2004 System Development Life Cycle Dengan Beberapa Pendekatan. *Jurnal FASILKOM* **1** 39-47

[12] Solichin A 2016 Pemrograman web dengan PHP dan MySQL (Jakarta: Penerbit Budi Luhur)

[13] Nugroho A 2009 rekayasa perangkat lunak menggunakan UML dan JAVA. (Bandung: Penerbit Andi)

[14] Han J, Pei J and Kamber M 2011 *Data mining: concepts and techniques* (New York: Elsevier)

[15] Suhendi H Y, Ramdhani M A and Irwansyah F S 2018 Verification Concept of Assessment for Physics Education Student Learning Outcome. *International Journal of Engineering & Technology (UEA)* **7** 321-325

[16] Pamoragung A, Suryadi K and Ramdhani M A 2006 Enhancing the Implementation of E-Government in Indonesia through the High-Quality of Virtual Community and Knowledge Portal Design

[17] Ramdhani M A 2013 *Metodologi Penelitian dalam Riset Teknologi Informasi* (Bandung: UIN Sunan Gunung Djati Bandung)

[18] Ramdhani M A, Maylawati D S and Amin A S and Aulawi H 2018 Requirements Elicitation in Software Engineering. *International Journal of Engineering & Technology (UEA)* **7** 772-775

[19] Aulawi H, Ramdhani M A, Slamet C, Ainissyifa H and Darmalaksana W 2017 Functional Need Analysis of Knowledge Portal Design in Higher Education Institution. *International Journal of Soft Computing* **12** 132-141

[20] Darmalaksana W, Ramdhani M A, Cahyana R and Amin A S 2018 Strategic Design of Information System Implementation at University. *International Journal of Engineering & Technology (UEA)* **7** 787-791

[21] Suryadi K and Ramdhani M A 2002 *Sistem Pendukung Keputusan* (Bandung: Remaja Rosdakarya)

[22] Maylawati D S and Ramdhani M A and Amin A S 2018 Tracing the Linkage of Several Unified Modelling Language Diagrams in Software Modelling Based on Best Practices. *International Journal of Engineering & Technology (UEA)* **7** 776-780

[23] Slamet C, Rahman A, Sutedi A, Darmalaksana W, Ramdhani M A and Maylawati D S 2018 Social Media-Based Identifier for Natural Disaster. In *IOP Conference Series: Materials Science and Engineering* **288** 1 012039

[24] Slamet C, Andriyan R, Maylawati D S A, Darmalaksana W and Ramdhani M A 2018 Web Scraping and Naïve Bayes Classification for Job Search Engine. In *IOP Conference Series: Materials Science and Engineering* **288** 1 012038

[25] Gerhana Y A, Zulfikar W B, Ramdani A H and Ramdhani M A 2018 Implementation of Nearest Neighbor using HSV to Identify Skin Disease. In *IOP Conference Series: Materials Science and Engineering* **288** 1 012153

[26] Rahman A, Slamet C, Darmalaksana W, Gerhana Y A and Ramdhani M A 2018 Expert System for Deciding a Solution of Mechanical Failure in a Car using Case-based Reasoning. In *IOP
Conference Series: Materials Science and Engineering 288 1 012011

[27] Slamet C, Rahman A, Ramdhani M A and Darmalaksana W 2016 Clustering the Verses of the Holy Qur'an using K-Means Algorithm. Asian Journal of Information Technology 15 24 5159-5162

[28] Zulfikar W B, Prasetyo, P K and Ramdhani M A 2018 Implementation of Mamdani Fuzzy Method in Employee Promotion System. In IOP Conference Series: Materials Science and Engineering 288 1 012147

[29] Maylawati D S A, Ramdhani M A, Rahman A and Darmalaksana W 2017 Incremental technique with set of frequent word item sets for mining large Indonesian text data. In Cyber and IT Service Management (CITSM), 2017 5th International Conference on 1-6

[30] Taofik A, Ismail N, Gerhana Y A, Komarujaman K and Ramdhani M A 2018 Design of Smart System to Detect Ripeness of Tomato and Chili with New Approach in Data Acquisition. In IOP Conference Series: Materials Science and Engineering 288 1 012018

[31] Witten I H, Frank E, Hall M A and Pal C J 2016 Data Mining: Practical machine learning tools and techniques (Burlington: Morgan Kaufmann)

[32] Marisa F 2013 Educational Data Mining (Konsep dan Penerapan). Jurnal Teknologi Informasi: Teori, Konsep, dan Implementasi 4 2 90-97

[33] Sartika D and Sensuse D I 2017 Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. Jurnal Teknik Informatika dan Sistem Informasi 3 2 151-161

[34] Rifai A 2017 Kajian Algoritma C4.5, Naive Bayes, Neural Network Dan SVM Dalam Penentuan Kelayakan Pemberian Kredit. Jurnal Sistem Informasi 5 2 176-182

[35] Astuti P 2016 Komparasi Penerapan Algoritma C45, KNN dan Neural Network Dalam Proses Kelayakan Penerimaan Kredit Kendaraan Bermotor. Faktor Exacta 9 1 87-101

Acknowledgments
The authors expressed the appreciation and thank you to the Research and Publishing Centre of UIN Sunan Gunung Djati Bandung, which has provided funding support for the publication of this article.