Prediction of Job Suitability of College Graduate Candidates Using Data Mining Algorithms

Vanessa Stefanny[1], Arief Wibowo[2]
{fannybataona@gmail.com[1], arief.wibowo@budiluhur.ac.id[2]}

STMIK Insan Pembangunan Tangerang, Indonesia[1],
Universitas Budi Luhur Jakarta, Indonesia[2]

Abstract. Large amount and volume of the database in an educational institution can be used to improve educational quality, for example is to know students‘ performance, who are expected to have appropriate work (relevant) with the study program that leads. This study aims to find the best method in predicting job suitability for college graduate candidates that used as one of the standard assessment of National Accreditation Board in Indonesia. The classification modeling was completed by applying the ICT competency standards in Indonesia known as SKKNI into the dataset. This research was conducted by comparing the result of decision tree J48,Naive Bayes,KNN (K=1) and Random Forest. The performance test of the algorithm using K-folds cross-validation method has shown that J48 is the best algorithm for this case with accuracy rate is 85%. It can be concluded that J-48 can be applied in designing prototypes.

Keywords: data mining, classification, J48 decision tree algorithm, job suitability, SKKNI.

1 Introduction

Data mining is one of the methods to analyze large datasets for getting information. The extracting data process to obtain information plays an essential role for management level in almost every field for decision support, including educational fields [1][2]. Many data in educational fields are available to be researched using data mining to improve the quality of education. Government as an authorized institution also makes an assessment related to suitability between job and final-year students' as a college graduate candidate which has become one of the key performance indicators of the institution.

Indonesian National Work Competency Standards or abbreviated SKKNI is a formulation of work skills that includes aspects of knowledge, skills, and expertise and work attitudes. SKKNI is structured according to the competence function in various fields so that the SKKNI of multiple sectors can be used as a reference in conducting competency-based education/training and in the implementation of competency test (competency certification).

In this research, the researcher applied data mining techniques to predict relevance between future job and final-year students' major based on each student's academic and general information referring to SKKNI. This comparative study was conducted by comparing the model that was built with two types of data, i.e., learning data with target classes determined by the Head of Study Program (without SKKNI validation), and learning data that has been

ICCSET 2018, October 25-26, Kudus, Indonesia

DOI 10.4108/eai.24-10-2018.2280576

validated by experts in determining the target class based on SKKNI. The classification modeling applied to the data mining techniques.

2 Related Works

Sudheep Elayidom and Sumam Mary Idikkula [3] researched by comparing 3 (three) data mining methods (Decision Tree, Naive Bayes, and Neural Networks) to test student data with the goal of helping students make career decisions using technologies such as data mining. The three models have given comparable performances based on accuracy, and one of the models can be used depending on the platform for implementation.

Elakia, Gayatri, Aarthi and Naaren J. [4] conducted a comparative research of 3 (three) decision tree models, i.e., ID3, CHAID and C-45 to justify data mining techniques which include classification used in educational databases to suggest career options for high school students and also to predict the potentially violent behaviour among students by adding extra parameters other than academic details. The research was evident that ID3 outperformed every other decision trees when all the performance measures were compared.

Mythili and Mohamed Savanas [5] concluded that random forest was the most accurate classifier and took less time to build the model than any other classifier. This research also found that the attendance, parent education, locality, gender, and economic status were the high potential parameters affecting student performance in the examination.

Research conducted by Nursalim, Suprapedi and Himawan [6] on the comparison of 5 (five) data mining algorithms which included Naïve Bayes, Decision Tree, Neural Network and Support Vector Machine, K-Nearest Neighbor to determine the graduates work field. The research was evident that K-Nearest Neighbor had better performance based on accuracy and AUC value.

Budanis Dwi Meilani and Nofi Susanti [7] used Naive Bayes to get the pattern of 220 student's graduation rates and reached 99.83% of accuracy. Fajrian Nur Adnan, Khaafiza Nuur Rakhmah, and Adhitya Nugraha [8] researched about making application in predicting the field of work based on value acquisition and built prediction model of decision tree method and case-based reasoning approach by utilizing graduates data from institution career center as data learning.

Mashael Al-Barak and Muna Al-Razgan [9] used J48 algorithm to predict students'final GPA and found that Java 1, Database principles, Software Engineering 1, Information Security, Computer ethics and Project 1 as mandatory courses which had an essential impact on final GPA. Vinaya Patil, Shiwani Suryawanshi, Mayur Saner and Viplav Patil [10] compared various classification data mining techniques, i.e., Naive Bayes, LibSVM, C4.5, random forest and ID3 to predict engineering students performance to lower dropout rates and terms of grade to improve the quality of engineering graduates.

This research would compare two learning datasets with modification the data learning model of graduates which had been explored based on SKKNI as target class classification standard. The purpose of this research was to find the best algorithm based on the accuracy and to implement the rules into a prototype of prediction of relevance between job and study programme based on SKKNI as one of the graduate standards in the institution.

3 Pre-processing

This research used 130 records of STMIK Masa Depan's graduates from 2010 until 2015. All general and academic information of students were stored in different single tables. The first phase of this research was to check and replace multivalue data into numeric data by counting value in the given attribute. After raw dataset was checked, the next step was to select some relevant properties for prediction in a database to be mined.

Some collected attributes had less efficiency for prediction unless being combined. The attributes like entry year of study and graduation year were merged into one form that contained information about the length of study time.

Modeling dataset must be labeled to indicate target class with classification. The attribute relevance status is applied to predict the relevance between future job and student's majors at STMIK Masa Depan which is classified into two classes (Relevant and Irrelevant) based on accreditation standard from National Accreditation Board for Universities or known as BAN-PT.

4 Building The Model

In this research, the preprocessed dataset would be processed by 4 (four) data mining algorithm, i.e., J-48, Naive Bayes, K-NN (K=1) and Random Forest. Comparative research would be done in the following steps.

4.1 Data Selection

The raw dataset for modeling had 130 records and 19 attributes. All of these attributes were collected and analyzed to get better result of classification and were reduced into 11 attributes and would be divided into 2 (two) sets, i.e., 100 records for modeling and 30 records as prototype testing data. Attributes on the processed datasets can be seen in Table 1.

Table 1. List of attributes.

Attributes	Description	Possible Values
Tgl_lhr	Date of Birth	Birth of date
JK	Sex	{P (Female), L(Male)}
Jurusan	Major	{Information System, Informatics}
Lama_Kuliah	Study time	{1,2,3,4,>4}
BBM	Number of BBM Scholarships earnings	{1,2,>2}
PPA	Number of PPA Scholarships earnings	{1,2,>2}
Organisasi	Number of Organizations	{1,2,>2}
Seminar	Amount of seminar participation	{1,2,>2}
MK_ulang	The number of failed courses	{1,2,>2}
IPK	GPA	{>2.0 , <4.0}
Status	Status of job relevance to student's major	{*Relevan* (Relevant), *Tidak Relevan* (Irrelevant)}

Table 1 shows that there are seven attributes used for modeling. Consists of 6 label classes and one attribute as the target class. Target classes are relevant and irrelevant.

4.2 Data Selection

This study used decision tree J-48, Naive Bayes, K-NN (K=1), Random Forest and would be compared and validated using 10-folds cross-validation technique that would generate accuracy, precision, recall, specificity and AUC [11]. The model which produced a better result would be implemented to build a prototype using Java. In this research, there were several steps of modeling phases, i.e.:

4.2.1 Crosschecking raw dataset

The first step of this study was modeling the dataset to obtain statistical value and re-check some inconsistency records in the dataset. All ambiguous data would be checked by an internal member of the institution to get a valid dataset.

4.2.2 Modeling phase

The classification process of relevance between job and major is having a connection to assessment by BAN-PT. The dataset would be verified by the head of each department at STMIK Masa Depan based on The BAN-PT Standard. The dataset which had been verified would be tested using WEKA tools. The result of the statistical value in this step is shown in Section V.

4.2.3 SKKNI learning

SKKNI implementation plan would be learned first from the experts to gather the required knowledge. Before doing classification, the researcher would gather all knowledge about SKKNI, i.e., the function of SKKNI, how to implement SKKNI and which SKKNI sector should be used. The knowledge could be found at the official site and be obtained from the experts. After gathering all of SKKNI knowledge, the first step to implement SKKNI in this research was SKKNI selection. This step needed verification of SKKNI expert to get a suitable list of SKKNI to be used in the dataset.

4.2.4 Re-label using SKKNI standard

Each list of SKKNI in different field areas of Information Technology and Communications which had been verified by experts was restudied before applying it to the classification process to guarantee the re-labeled process. This process was approved by the institution to be conducted for research to obtain a better result.

4.2.5 Validate the classification process

Re-labeling process would change the amount of Relevant and Irrelevant class from the previous dataset. Relabeling values of target attribute in the classification process were done by checking each graduate's job in the database to be matched with suitable SKKNI. Before modeling the

SKKNI version of the dataset, all of the target values must be validated by SKKNI experts to ensure the classification.

4.2.6 Re-modeling with SKKNI validation by the expert

SKKNI version of the dataset which had been relabeled and validated by the SKKNI experts would be tested by decision tree J-48, Naive Bayes, K-NN (K=1) and Random Forest using WEKA tools. The result of modeling is shown in Section V.

4.2.7 Evaluation Phase

This research would be evaluated based on accuracy, precision, recall, specificity and AUC value. The generated rules from modeling would be implemented into a prototype which would be tested using a dataset for prototype verification (30 records).

5 Results and Discussion

Before applying SKKNI as the standard of the classification process, the old version of the dataset which used internal institute standard would be tested using J-48, Naive Bayes, K-NN (K=1) and Random Forest to show the accuracy of prediction process before doing modification in standard classification. The results would be calculated to get the performance result in Table 2.

Table 2. Performance test result (dataset without SKKNI validation).

Method	Accuracy	Precision	Recall	Specificity
J-48	68%	72,84%	85,51%	29,03%
Naive Bayes	70%	80%	75,36%	58,06%
K-NN (K=1)	56%	71,19%	60,87%	45,16%
Random Forest	76%	74,19%	100%	22,58%

Table 2 has shown that the Random Forest algorithm is the best classification method for modeling without including SKKNI validation by the expert in the learning data. As the purpose of this research was to test the SKKNI dataset as a classification standard, the modeling phase would be repeated using the dataset which is validated by the expert of SKKNI in re-classing.

Table 3. Performance test result (dataset with SKKNI validation).

Method	Accuracy	Precision	Recall	Specificity
J-48	85%	88,16%	91,78%	66,67%
Naive Bayes	76%	85,51%	80,82%	62,96%
K-NN (K=1)	67%	73,81%	84,93%	18,52%
Random Forest	78%	76,84%	100%	18,52%

Table 3 shows the performance test result of SKKNI dataset using 4 (four) data mining algorithms. The best algorithm is the Decision Tree (J-48) with an accuracy value of 85%. The result proved that in its accuracy, Decision Tree J-48 got better performance among other algorithms and the generated rules could be implemented (Fig. 2) into a prototype of Prediction Relevance Between Future Job and Final-Year Student's Major in STMIK Masa Depan.

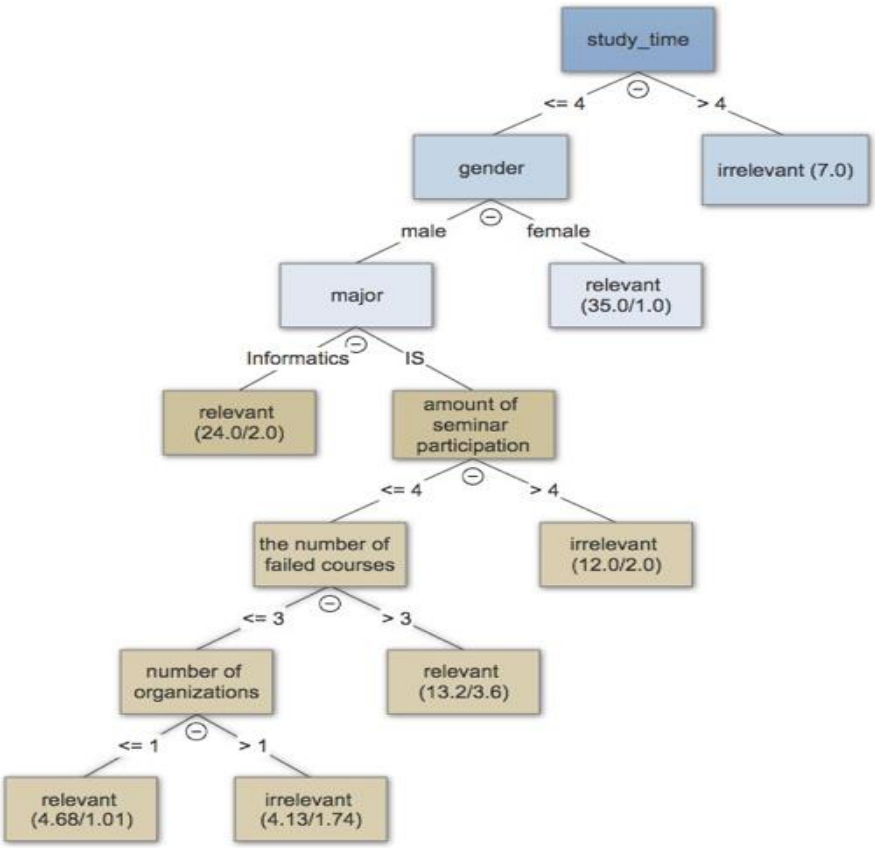

Fig. 2. The generated J-48 tree for prediction relevance based on SKKNI between job and student's major.

Figure 2 showed that attribute of *lama_kuliah* (study_time) is the root node of the tree and followed by other attributes. This value is derived from the process of subtraction the year of graduation with the year of entry. The tree was transformed into rules using programming language. This research transformed the generated tree into a prototype using JAVA and MySQL as the database management system. A prototype of a prediction system built on the decision tree model is shown in Figure 3.

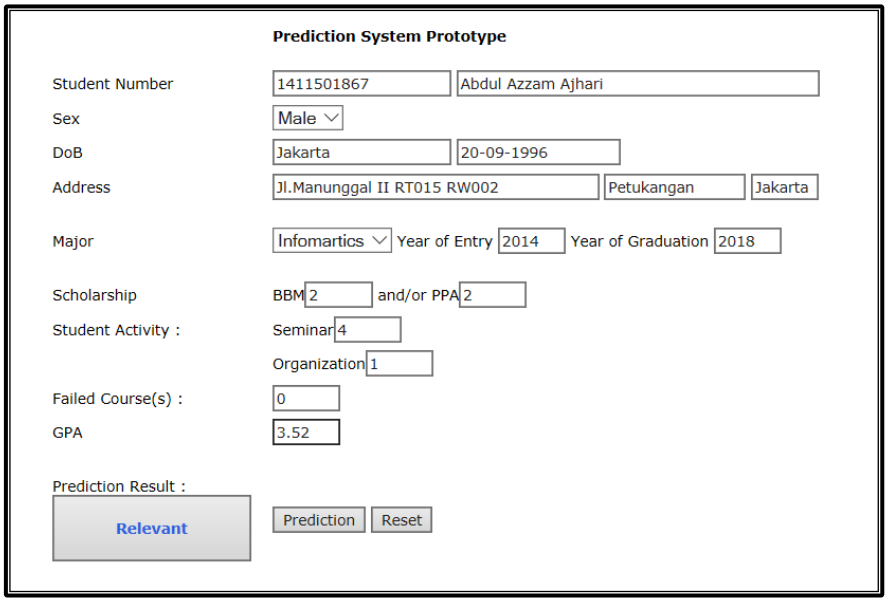

Fig. 3. The GUI design of system prototype.

Figure 3 shown the GUI form to process only one piece of student information that was input manually and then, after pressing the Predict button, the result automatically indicates the relevance between future job and graduates major. The result would show 'Relevant' (*Relevan*) and 'Irrelevant' (*Tidak Relevan*) based on the generated rules (Fig. 2).

6 Conclusion

Data mining is used to find a pattern of information in large data. The results of this comparative study concluded that by accuracy value, Decision Tree J48 is the best method among other algorithms. The process of data learning has involved the opinion of SKKNI experts produce models with an accuracy value of 85%. This result is better if compared to the learning process does not involve SKKNI experts in data learning. Based on these results, it can be developed a prototype system that can predict the relevance of future work with student data during the study.

Acknowledgments

The researcher highly appreciated STMIK Masa Depan who has permitted gathering graduates database for this research. The researcher also highly appreciated SKKNI experts who have helped to share knowledge, validate and verify the results based on SKKNI.

References

[1] D. A. Alhammadi and M. S. Aksoy: Data Mining in Education- An Experimental Study. Int. J.

Comput. Appl., vol. 62, no. 15, pp. 31–34 (2013)

[2] R. V. Monika Goyal: Applications of Data Mining in Higher Education. Int. J. Comput. Sci. Issues, vol. 9, no. 2, pp. 113–120 (2012)

[3] S. Elayidom, S. M. Idikkula, and J. Alexander: A Generalized Data Mining Framework for Placement Chance Prediction Problems. Int. J. Comput. Appl., vol. 31, no. 3, pp. 40–47 (2011)

[4] Elakia, Gayathri, Aarthi, and N. J.: Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students. Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 3, pp. 4649–4652 (2014)

[5] M. M.S. Mythili and D. A.R.Mohamed Shanavas: An Analysis of Students' Performance using Classification Algorithms. IOSR J. Comput. Eng., vol. 16, no. 1, pp. 63–69 (2014)

[6] Nursalim, Suprapedi, and H. Himawan: *Klasifikasi Bidang Kerja Lulusan Menggunakan Algoritma K-Nearest Neighbor.* J. Teknol. Inf., vol. 10, no. 1, pp. 31–43 (2014)

[7] B. D. Meilani and N. Susanti: *Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naïve Bayes.* J. Ilm. NERO, vol. 1, no. 3, pp. 182–189 (2015)

[8] F. N. Adnan, K. N. Rakhmah, and A. Nugraha: *Aplikasi Berbasis Sistem Pakar Untuk Memprediksi Peluang Kerja Calon Lulusan Mahasiswa Sistem Informasi Universitas Dian Nuswantoro*. J. Syst. Inf., pp. 27–38 (2016)

[9] M. A. Al-Barrak and M. Al-Razgan: Predicting Students Final GPA Using Decision Trees: A Case Study. Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533 (2016)

[10] V. Patil, S. Suryawanshi, M. Saner, and V. Patil: Student Performance Prediction using Classification Data mining Techniques. Int. J. Res. Emerg. Sci. Technol., vol. 4, no. 4, pp. 15–18 (2017)