An Internet-based Student Admission Screening System utilizing Data Mining

Dolluck Phongphanich and Wirat Choonui
Department of Science and Technology,
Suratthani Rajabhat University,
Suratthni, Thailand

***Abstract*—This study aimed to propose an internet-based student admission screening system utilizing data mining in order for officers to reduce time to evaluate applicants as well as for the faculty to use less human resources on screening applicants that meets their proficiency and criteria of each department. Another benefit is that the system can help applicants efficiently choose a specialization that is suitable to their proficiency and capability. The system used a decision tree based classification method. Prior to system development, six models were created and tested to find the most efficient model which would later be applied for development of internet-based student admission screening system. The first three of six models employed a k-fold cross validation technique, while the remaining three models use a percentage split test technique. Experiment results revealed that the most efficient model was the data classification model that uses Percentage Split (80), which provided the precision of 87.90%, recall of 87.80%, F-measure of 87.60% and accuracy of 87.82%. To make the efficient student admission screening system, this experiment selected a data classification model that implements Percentage Split (80).**

Keywords*—Classification method; data mining; decision tree; student admission screening

I. Introduction

Undergraduate student admission of educational institutions in Thailand is crucial because it directly affects to education management, budget planning for institution administration and education management, and lastly educational quality and standard indicator of each university that mainly concentrates on students. The efficient student admission as well as nurturing students throughout their enrolled curriculum until they complete the study in high quality under a specified timeframe are therefore what the institutions realize and pay attention to [1]. Faculty of Science and Technology, Suratthani Rajabhat University continuously receives a lot of applications and new students can enroll to the faculty in various ways. Each academic year, the university has to advertise itself in different ways, such as a roadshow and direct admission at high schools, billboard advertising, admission advertising via radio and newspapers, so as to gain a huge volume of applications, and this gives the institution more opportunity to get a number of candidates with appropriate knowledge and capabilities for further examination to finally select those candidates as new students of the university.[1]

Nonetheless, each student admission requires a number of personnel to evaluate student's profile so as to screen the right applicants given each department's criteria. And since criteria are different from one department to another, each student admission screening takes time and sometimes the screening does not serve unqualified students in accordance with a department's criteria, due to the fact that staffs evaluating those applicants are not from the department where students apply for. This results in maintaining a student status for an entire curriculum. That is, students are unable to complete the program or even finally drop out from studying.

This study aimed at developing an internet-based student admission screening system utilizing data mining to help reduce time as well as a number of personnel for evaluating applicants to select ones in accordance with their capability and criteria specified by each department. Besides, this system would help applicants choose the right specialization conforming to their proficiency and capability. The system was developed by analyzing student profiles to create six decision models for a decision tree based classification method, which is efficient and one of popular techniques for data mining. Those six models came from different modeling techniques: the first three models used a k-fold cross validation technique, while the next three models implemented a percentage split test technique. All models were then compared to each other to select the most efficient model for developing an internet-based student admission screening system. This student admission screening system will not only help save time and human resources on application screening, but also help applicants decide to select a specialization for studying which most fits with their characteristics and the university's objectives.

The next topics will describe related literature, research methodology, discussion of findings, and conclusion, respectively.

II. Literature Review

Sumitra Nuanmeesri develops an information system to

The Research and Development Institute of the Suratthani Rajabhat University (*sponsors*).

[1] Suratthani Rajabhat University, Department of Computer Science, "Recruitment Regulations for Students Admission", 2016, [Online] Available: http://www.sci.sru.ac.th/qts/devop.php.

forecast student admission via the internet with the aim of correctly and accurately forecasting student admission. As part of research methodology, the researcher creates and tests seven forecast models.[2] Three of those models use a k-fold cross validation technique, while the next three of the models employ a percentage split technique, and the last one apply a technique of separating data for training and testing a model. From the experiment, the technique of separating data for training and testing a model serves better performance on forecasting students than any other modeling technique as the former has the accuracy of 94%, precision of 94.30%, recall of 94.00% and F-measure of 93.70%. Decision tree classification rules underlying the most efficient model are utilized as part of development of information system to forecast student admission via the internet. The system is then evaluated by two sample groups comprising of experts (4 persons) and personnel (40 persons) based on mean and standard deviation. System performance evaluation shows that the average of experts was 4.17 while the average of personnel was 4.34. It can be concluded that the information system has satisfactory performance and can be applied to forecast student admission.

Supatkul Phakkachokh [2] applies data mining techniques to develop a model for selecting high school program with the objectives of discovering factors influencing selection of study program as well as capability to complete the chosen program successfully by using data mining techniques. Data used in this study is from study result of each subject and questionnaire on study program selection of high school students. A sample group consists of 850 students of Satri Si Suriyothai School enrolled in academic year 2012. The result shows that a high school study program selection model can represent what factors influence on study program selection and provide the accuracy of study program suggestion of 79.76%. It can be conclude from the model that a score of junior high school's basic subjects, including Thai language, mathematics, science, social studies, religion and culture and English language, as well as grade point average (GPA) are factors directly affecting to study program selection and success in completing the chosen program.

Raywadee Sakdulyatham adapts data mining techniques in knowledge based creation for education achievement prediction of Ratchaphruek College students to predict the right specialization so that academic advisors to use derived rules for providing academic advices.[3] Data used for modeling includes personal details and registration data of students from all of four specializations under Faculty of Business Administration, including Marketing, Business Computer, Management and Hotel and Tourism Management. The outcome is a model for analyzing student learning behaviors in each department which suggests that a study result of core finance subject group impacts to study result of restricted elective subject groups of Business Computer and Hotel and Tourism Management most, whereas a study result of core business subject group impacts to a study result of restricted elective subject groups of Marketing and Management most. Apart from that, a study result forecasting model was created for each specialization. The prediction model of study result for Business Computer has an accuracy of 73.49%, model for Marketing has an accuracy of 83.58%, model for Management has an accuracy of 78.12% and model for Hotel and Tourism Management has an accuracy of 86.67%.

Utcharaporn Juthapart, Kant Charoenjit and Phayung Meesad [1] adapt data mining techniques for providing suggestions of specialization selection to students, since most of students lack knowledge, understanding and experience about choosing a specialization, so they decide to pursue the inappropriate one. The technique adopted in this study is a decision tree algorithm, which is similar to that of Sumitra Nuanmeesri. Both researchers categorize grades into three groups: High (grade A, B+ and B), Medium (grade C+ and C) and Low (grade D+, D and F). Findings revealed that using a decision tree algorithm to categorize students of all specializations is very efficient as all models have an accuracy of more than 80%.

Teerapong Sungsri [3] applies the concept of data mining for analyzing candidates' profile and then stores the analysis result in a database for planning of future student admissions. The research comprises of two modules. The first module is for analysis of specialization selection behavior by using a simple k-means clustering technique, which results in four behavioral groups. The second module is for searching for association rules among groups of applicant behaviors by applying an Apriori algorithm with a confidence of 0.9. The second module is for comparing two models forecasting a number of new students. One model is created by a decision tree algorithm which is similar to Utcharaporn Juthapart, Kant Charoenjit and Phayung Meesad [1] and Sumitra Nuanmeesri with accuracy of 93.76%, while the other model is created by a multilayer perception-based artificial neural network model with accuracy of 93.60%.

From all related researches aforementioned, a classification technique, which is one of data mining techniques currently popular, is applied on educational data with the use of decision tree algorithm for modeling. Although this study applies a decision tree based classification techniques like related literature, but this study is differentiated from the others in a way to create a model and objectives of utilizing data from a model to develop an internet-based student admission screening system to facilitate related personnel as well as to help applicants make a decision on selecting a specialization appropriate to their proficiency. The next section will describe research methodology.

III. METHODOLOGY

A methodology of this research was divided into two stages. The first stage will be data analysis using data mining and the second stage will be development of internet-based student admission screening system. Both stages will be presented in the following sections.

[2] S. Nuanmeesri, "Developing Information System to Forecast the Student Admission via the Internet". Suan Sunandha Rajabhat University (In Thai), 2012. [Online]. Available: http://www.eresearch.ssru.ac.th/bitstream/123456789/330/1/ird_036_55%20%281%29.pdf.

[3] R. Sakdulyatham, "Utilizing Data Mining Techniques in Knowledge Based Creation for Education Achievement Prediction of Ratchaphruek College Students. (In Thai), 2009. [Online]. Available: http://www.rpu.ac.th/ebook/54/54-4.pdf (2009).

A. Data analysis using data mining

We followed Cross-Industry Standard Process for Data Mining (CRISP-DM) (shown in Fig. 1) which has six phases as follows:

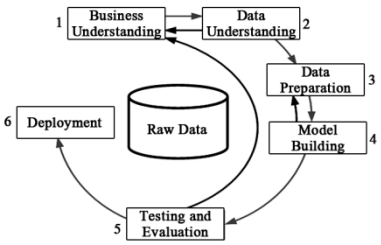

Fig. 1. Cross-industry standard process for data mining [4].

1) Business understanding and data understanding

The first and second phase of CRISP-DM is business understanding and data understanding, respectively. For business understanding, we targeted that this experiment helps facilitate personnel on quickly screening applicants regarding to criteria defined by Faculty of Science and Technology, helps reduce a number of personnel for evaluating candidate qualifications, aids students on choosing a specialization that meets their proficiency, thereby reducing student dropouts, as well as helps planning for future student admissions. And in terms of data understanding, we studied data files managed by Office of the Registrar by looking into data characteristics and validating not only data integrity, but also possibilities of using data for analysis.

2) Data preparation

The third phase is about data preparation which covers activities to improve data quality prior to analysis by using a decision tree algorithm. Those activities include verification of data integrity and completeness, data cleaning which includes feature verifications in terms of missing value, noisy data, errors and outliers, as well as data inconsistencies. Data preparation of this research can be explained below.

At the beginning, data collection was performed on 984 data files of new students of each specialization of Faculty of Science and Technology, Suratthani Rajabhat University, including a student ID, full name, national ID, address, contact telephone number, highest level of education, selected specialization for an undergraduate study, score from each subject taken in a high school education, such as mathematics, science and English language, and GPA per semester. The data files were during academic year 2010-2012.

The next step was to perform data cleaning and preparation by removing some features in the sample to keep only necessary features for further analysis, which in this study included a highest level of education, selected specialization for an undergraduate study, score from each subject taken in a high school education, such as mathematics, science and English language, and GPA per semester. For a study result of main subjects, an average score of each subject group would be determined by considering a score for five semesters. For example, a mathematics subject group 1st – 5th semester was 6 periods a decision tree algorithm and student admission criteria set by Faculty of Science and Technology, Suratthani Rajabhat University.

For the last step of data preparation, data would be transformed to a proper format for further analysis by applying a decision tree algorithm on continuous and discrete quantitative data; for instance, a score of each subject and average score across five semesters are continuous data, so to prepare data for data mining, the quantitative data had to be transformed to a nominal scale as presented in Table 1. To illustrate, suppose that a student gets a score within 0.00-0.90, a nominal value will be F, meaning that the student fails to pass the criteria. For a score within 1.00-1.49, a nominal value will be T, meaning that the student's score is terrible. For a score within 1.50-1.99, a nominal value will be L, meaning that the student's score is low. For a score within 2.00-2.49, a nominal value will be M, meaning that the student's score is medium. For a score within 2.50-2.99, a nominal value will be G, meaning that the student's score is good. Lastly, for a score within 3.00-4.00, a nominal value will be E, meaning that the student is excellent, respectively.

TABLE. I. SPECIFIES A VALUE IN EACH SCORE SCALE

Score Scale	Value
0.00-0.90	Fail (F)
1.00-1.49	Terrible (T)
1.50-1.99	Low (L)
2.00-2.49	Medium (M)
2.50-2.99	Good (G)
3.00-4.00	Excellent (E)

3) Modeling

The fourth phase is model creation or so called modeling. An algorithm used to analyze a sample to build a model is J48 decision tree classifier. J48 decision tree is one of the decision tree families that can construct a tree for the purpose of improving prediction accuracy and produce both decision tree and rule-sets; The J48 decision tree classifier is among the most popular and powerful decision tree classifiers [5]. In this phase, we created six models. Model 1 to Model 3 used a k-fold cross validation technique, which partitions a sample into k equal sized sub samples. The first subsample is retained for testing a model, while the remaining k – 1 sub samples are used as training data. The cross-validation process is then repeated k times (folds). Model 4 to Model 6 used a percentage split test technique, which separates data into two parts: the first part is for testing while the second part is for training. More specifically, Model 1 partitioned data into five sub-samples equally, kept the first sub sample for testing and left 2nd – 5th sub-sample for training, and then repeated for five folds. Similarly, Model 2 partitioned data into 10 sub-samples equally and Model 3 partitioned data into 100 sub-samples equally. Model 4 slated the sample in 70:30 ratios, meaning 70% of an original sample was used for training whereas the remaining 30% of an original sample was used for testing. Similar to Model 4, Model 5 slated the sample in 80:20 ratios and Model 6 slated the sample in 90:10 ratios. WEKA software was used to develop models, and after that those models were compared to each other in the area of accuracy, precision, recall and F-measure to find the most efficient one for development of internet-based student admission screening

system which will be described in the next stage. Performance of each of six models was presented in Table 2.

TABLE. II. PRESENTS PERFORMANCE OF EACH MODEL

Modeling	Times(Seconds)	Precision (%)	Recall (%)	F-Measure(%)	Accuracy (%)
Cross validation (5 folds)	0.11	88.00	88.50	87.40	87.50
Cross validation (10 folds)	0.02	87.90	87.40	87.30	87.40
Cross validation (100 folds)	0.02	88.10	87.60	87.50	87.60
Percentage Split (90)	0.00	88.50	87.80	87.50	87.76
Percentage Split (80)	0.10	87.90	87.80	87.60	87.82
Percentage Split (70)	0.00	87.40	87.10	86.90	87.12

4) Testing and evaluation

The fifth phase is about testing and evaluation of generated models to see the efficiency, error and level of accuracy of each model so as to get the right model for real usage. Evaluation is measured in terms of precision, recall, F-measure and accuracy. In this stage, the accuracy of each model is compared to that of other models to find the most efficient one. All measures can be derived from a confusion matrix as presented in Fig. 2 and calculated by using below formulas:

		Predicted Class	
		Class=Yes	Class=No
Actual Class	Class= Yes	True Positive (TP)	False Negative (FN)
	Class= No	False Positive (FP)	True Negative (TN)

Fig. 2. Shows a confusion matrix.

1) Precision of a particular model can be measured by considering each class [6].

$$Precision = TP/(TP + FP) \quad (1)$$

2) Recall of a particular model can be measured by considering each class.

$$Recall = TP/(TP + FN) \quad (2)$$

3) F-measure is a measurement of precision and recall at the same time for a particular model by considering each class [5].

$$F\text{-}measure = (2 \times Precision \times Recall)/(Precision + Recall) \quad (3)$$

4) Accuracy is a measurement of integrity of a particular model by considering every class [5].

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (4)$$

5) Deployment

The sixth phase is an application of research findings. In this study, the most efficient model selected as an input of the next stage was a model implemented Percentage Split (80).

B. Development of internet-based student admission screening system utilizing data mining

The system was a web application based on PHP programming language and underlying database run on MySQL. The data analysis result from data mining in the first section was used together with student admission criteria of Faculty of Science and Technology's curriculums, which conforms to the university's targets. The system has two main functions. The first main function is for general users who can use the system to help suggest a specialization provided by the faculty according to their proficiency. This suggestion can be used to support decision making on applying an undergraduate program. The second main function is for personnel who can do basic screening from applicants' profile to see whether they pass the faculty's criteria by inputting a profile of each applicant or importing multiple applicants at once. Details of all functionalities will be further discussed in results and discussion section below.

IV. RESULTS AND DISCUSSION

In this study, we will present the results in two sections. The first section is about data analysis using data mining techniques and the second section is about development of student admission screening system that utilize data mining techniques.

A. Data analysis by using data mining techniques

In this section, modeling was done by using decision tree methods. All six models were compared in terms of accuracy, precision, recall and F-measure. A performance comparison was shown in Fig. 3.

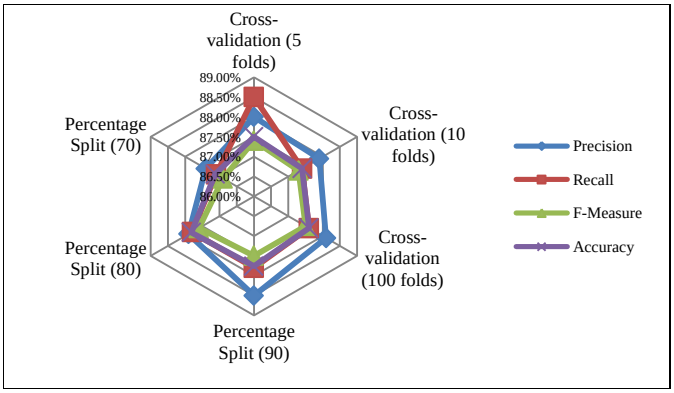

Fig. 3. Graphically presents performance of each model.

From an experiment of classification by using six models implementing different techniques, it was found from a comparison on precision, recall, F-measure and accuracy of all six models that all models gave the similar result shown in Table 3.

TABLE. III. PERFORMANCE OF EACH MODEL

Modeling	Order	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
Cross-validation (5 folds)	4	88.00 (3)	88.50 (1)	87.40 (4)	87.50 (4)
Cross-validation (10 folds)	5	87.90 (4)	87.40 (5)	87.30 (5)	87.40 (5)
Cross-validation (100 folds)	3	88.10 (2)	87.60 (4)	87.50 (2)	87.60 (3)
Percentage Split (90)	2	88.50 (1)	87.80 (2)	87.50 (2)	87.76 (2)
Percentage Split (80)	1	87.90 (4)	87.80 (2)	87.60 (1)	87.82 (1)
Percentage Split (70)	6	87.40 (6)	87.10 (6)	86.90 (6)	87.12 (6)

When considering each aspect, it was found from the experiment that the model with highest precision (represented in %) was 90 Percentage Split (88.50), followed by Cross validation (100 folds) (88.10), Cross validation (5 folds) (88.00), Cross validation (10 folds) and Percentage Split (80) (both at 87.90), and lastly Percentage Split (70) (87.40).

In terms of recall (represented in %), the experiment revealed that a model with highest recall was Cross validation (5 folds) (88.50), followed by Percentage Split (90) and Percentage Split (80) (both at 87.80), Cross validation (100 folds) (87.60), Cross validation (10 folds) (87.40), and Percentage Split (70) (87.10), respectively.

Next, for F-measure (represented in %), the model with highest F-measure was Percentage Split (80) (87.60), followed by Cross validation (100 folds) which gets the same F-measure as Percentage Split (90) (87.50), Cross validation (5 folds) (87.40), Cross validation (10 folds) (87.30), and lastly Percentage Split (70) (86.90).

Finally, as per accuracy (represented in %), the model with highest accuracy was Percentage Split (80) (87.82), followed by Percentage Split (90) (87.76), Cross validation (100 folds) (87.60), Cross validation (5 folds) (87.50), Cross validation (10 folds) 87.40, and Percentage Split (70) (87.12).

B. *Development of internet-based student admission screening system utilizing data mining*

They are of two types: component heads and text heads. This section will present user interfaces of the internet-based student admission screening system, which comprises of two sections.

1) General user section.
2) Officer section.

1) General user section

General users can use this internet-based student admission screening system to know a guideline in selecting a specialization respecting to their proficiency and capability. To tailor the guideline for users, the system will predict from a level of study results from a high school education based on a model from data mining and admission criteria of the faculty, which conform to the university targets. An input screen for general users is shown in Fig. 4.

A screen in Fig. 4 is for general users to input study results, which will be used by the system to suggest a specialization based on the student's proficiency. To do so, an applicant has to choose (subject) a designated specialization in Preferred Specialization (in case of not preferring any specializations, the system will perform analysis for all specializations) and select an education background. Then, the user has to fill in a score of each subject group for each of four semesters and average score of each subject group for all four semesters in the fifth semester row.

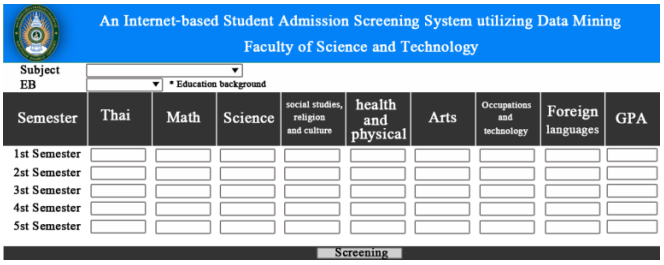

Fig. 4. An input screen for general users.

The most important data for screening is GPA of the fifth semester or GPA of the recent semester, in which the user is required to specify. After the user fills then click Screening, the system will display a screening result as shown in Fig. 5.

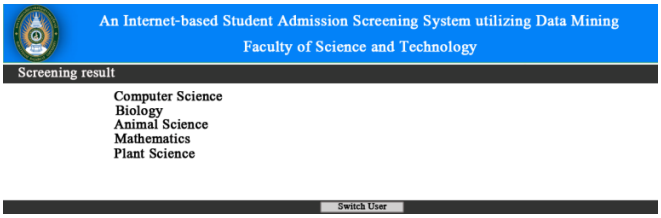

Fig. 5. Screening result.

2) Officer section

Officer section can use this internet-based student admission screening system to evaluate applicants as well as for the faculty to use less human resources on screening applicants that meets their proficiency and criteria of each department. The officer section was shown in Fig. 6. Three main menus for officers include:

1) Additional Subject.
2) Individual Screening.
3) Importing a CSV File.

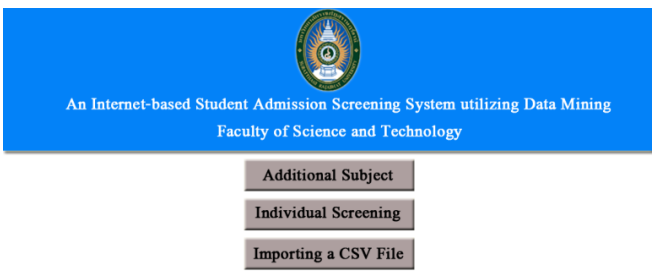

Fig. 6. Main menu for officers.

a) Additional Subject: Additional Subject is for searching and adding additional subjects in case that users would like to collect information of additional subjects from students. More specifically, the user can add and record additional information by using a screen shown in Fig. 7.

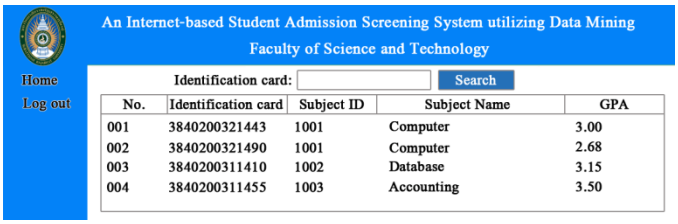

Fig. 7. An additional subject searching screen.

b) Individual Screening: Individual Screening is for officers to record a student profile which comprises of personal details; educational background and address as shown as an example in Fig. 8.

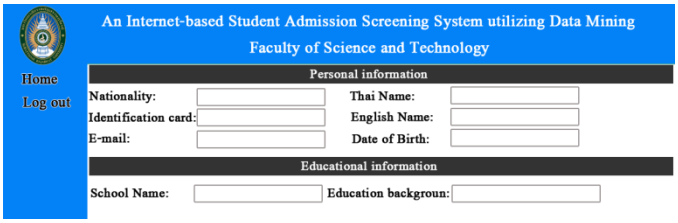

Fig. 8. An input screen to record a student profile for screening.

In terms of educational background, a study result is collected based on subject groups, including Thai language; mathematics; science; social studies, religion and culture; health and physical education; arts; occupations and technology and foreign languages and GPA. In case that a school or college does not provide ones of subject groups, an average grade of those subject groups can be blank. Indeed, the most important piece for screening is GPA of 5th semester or GPA of the recent semester, which will be retrieved by the system from the educational background section.

c) Importing a CSV File: Importing a CSV File is for officers to do screening of multiple applicants at once by inputting a number of student profiles. To use this function, an officer has to convert data into a CSV file with the condition that the CSV file must have a record with a given format as shown in Fig. 9.

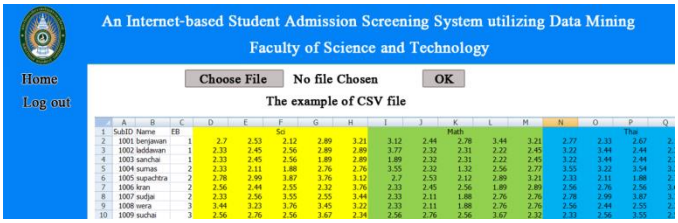

Fig. 9. A CSV file import screen.

V. Conclusion

The development of internet-based student admission screening system utilizing data mining used J48 decision tree to create a model through WEKA software, and then used the result for system development. 984 data sets of undergraduate applicants of Faculty of Science and Technology, Suratthani Rajabhat University during academic year 2010-2012 were used as samples for analysis. The system would help officers reduce time to do screening, help the faculty use less personnel for a screening process to find the right ones corresponding to their proficiency as well as criteria set by each department, leading to receive new qualified students aligning with a target group of each department. Moreover, this system would help applicants efficiently choose the right specialization according to their proficiency and capability. In addition, data from analysis can be used to make a decision on education management and budget planning for institution administration and learning management.

In this section, summary of this research will be described in the first section and suggestions will be discussed in the second section below.

A. Conclusion and discussion

This student admission screening system has a limitation of bring a model up to date, since it is developed based on applicants of academic year 2010-2012. If in the future some subjects are changed or added to in a given curriculum, the result from generated models may be incorrect.

B. Suggestion

1) Parameter settings in WEKA software results in different generations of classification rules or models; therefore, researchers should fine tune settings and use a large amount and variety of data for training and testing models in order to increase integrity and accuracy of screening.

2) This research idea can be improved and applied to similar works or used by another faculty, since capabilities of storing student profiles from the screening system, as well as recording a grade and additional subjects of applicants are in place.

C. Future work

In future work, we may adopt other data-mining techniques, such as anomaly detection or classification-based association, to gain more knowledge of the undergraduate applicants in Faculty of Science and Technology. We also plan to use data sets of undergraduate applicants from all departments of Suratthani Rajabhat University and compare the results with the data set from Faculty of Science and Technology.

Acknowledgment

We gratefully acknowledge financial support from the Research and Development Institute of the Suratthani Rajabhat University. We would like to express thanks to the Office of Academic Promotion and Registration as well as Department of Science and Technology, SRU for giving the information in this research and highly appreciates Dr. Nara Phongphanich, lecturer in Maejo University, Thailand for providing us advice during this study.

REFERENCES

[1] U. Juthapart, K. Charoenjit and P. Meesad, "Using Data Mining Technique to Selecting Majors for Students at the Faculty Information Technology Phetchaburi Rajabhat University". Joint Conference on ACTIS & NCOBA 2015, Jan 30-31, Nakhon Phanom, Thailand. ISSN: 1906-9006.

[2] S. Phakkachokh, "A Model for Selecting High School Program by Considering the Primary Subject Records Using Data Minig Techniques", Master's Thesis, Department of Science in Web Engineering, Faculty of Information Technology, Dhurakij Pundit University, 2013.

[3] T. Sungsri, "The Behavior Analysis on the Applying Major Selection and the Comparison of Model to Forecast the Numbers of New Students Using Data Mining Technique". The Tenth National Conference on Computing and Information Technology (In Thai), NCCIT2014 :pp.963-968, 2015.

[4] D. L. Olson and D. Denlen, "Advanced Data Mining Techniques., Springer-Verlag". ISBN 978-3-540-76916-3, 2008.

[5] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann publishers, San Francisco: CA, 2006.

[6] L. H. Witten, E. Frank and M. A. Hall, "Data Mining Practical Machine Learning Tools and Techniques", 3nd ed., Burlington, USA: Morgan Kaufmann publishers, 2011.