JURNAL ONLINE INFORMATIKA

JOIN (Jurnal Online Informatika)
Volume 6 No.1 | June 2021: 11-16
DOI: 10.15575/join.v6i1.608

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Comparison of C4.5 Algorithm and Support Vector Machine in Predicting the Student Graduation Timeliness

Agus Mailana[1], Andi Agung Putra[2], Sarifudlin Hidayat[3], Arief Wibowo[4]
[1,2,3]Department of Informatics, STAI Al-Hidayah Bogor, Indonesia
[4]Master of Computer Science, Universitas Budi Luhur

Article Info

Article history:

Received July 20, 2020
Revised March 26, 2021
Accepted May 18, 2021
Published June 15, 2021

Keywords:

college student
c4.5 Algorithm
support vector machine
timely graduation

ABSTRACT

In higher educational institutions, graduation rates are one of the many aspects to assess the quality of the learning process. Al-Hidayah Islamic University in Bogor is one of the established private Islamic universities to create skilled human resources with moral values required by many companies nowadays. Having another institution in Bogor as a competitor with the same direction and objective is a challenge for Al-Hidayah Islamic University. Thus a solution is required to face the competition. One solution is to predict the student graduation timeliness of the students using data mining method with classification function. The implemented methodology in the data mining is Discovery Knowledge of Database (KDD), starting from selecting, preprocessing, transformation, data mining, and evaluation/ interpretation. There were two Algorithm models used in this paper, namely C4.5 and Support Vector Machine (SVM). The classification procedure consists of predictor variables and one of the target variables. Predictor variables are gender, Grade Point Average, marital status, and job status. Rapid Miner software was used to process the data. The final results of both Algorithms show an 81% precision rate and 80% accuracy level for the C4.5 Algorithm, while SVM has an 88% precision rate and 85% accuracy level.

Corresponding Author:

Agus Mailana,
Department of Informatics,
STAI Al-Hidayah Bogor,
Jl. Raya Dramaga No.29, RT.03/RW.02, Margajaya, Kec. Bogor Bar., Kota Bogor, Jawa Barat 16116
Email: agus.mailana@gmail.com

1. INTRODUCTION

According to the Internal Quality Assurance Standard of higher educational institutions, timely graduation is one of the significant aspects to assess the quality of the institution. Al-Hidayah Islamic University Bogor applies the standard of Grade Point Average (GPA) of 3.0. Currently, the data stored in their database has not been more utilized to obtain knowledge that can be used in decision making. Since the number of Al-Hidayah Islamic University and the GPA standard have increased, research is necessary to predict the scores and the timeliness of student graduation in the present based on the existing data from the past.

There are many methods to predict the timeliness of student graduation, one of them is data mining. Using data mining as a database and statistic-based analysis technique is suitable to dig various potential information in an academic database.

The previous researches on student graduation timeliness prediction that have used data mining are PSO-based SVM [1], C4.5 Algorithm in Informatics Department [2], Support Vector Machine (SVM) with several scenarios [3], SVM Algorithm Optimization with PSO in LP3I Polytechnic Depok, Jakarta [4], research in Muhammadiyah University Ponorogo, class of 2012/2013, that used C4.5 [5], a combined method of Decision Tree and Artificial Neural Network [6], Decision Tree implementation using *Weka* application [7], Predicting the Student Graduation Rate to Determine the College Marketing Strategy Using Decision Tree [8], using Decision Tree C4.5 Algorithm with pruning technique [9], The Implementation of Decision Tree to Analyze the Resignation Possibility of Prospective New Students [10]. The C4.5 and Support Vector Machine (SVM) data mining techniques have been frequently used to predict student graduation timeliness. Therefore, this paper aims to compare both methods to see the difference in the accuracy levels resulting from both techniques by adding job status and marital status as variables in Al-Hidayah University Bogor.

2. METHOD

This study used Knowledge Discovery in Database (KDD), the stages of process model as shown in Figure 1. The explanations of each step of the research are as shown in figure 11:

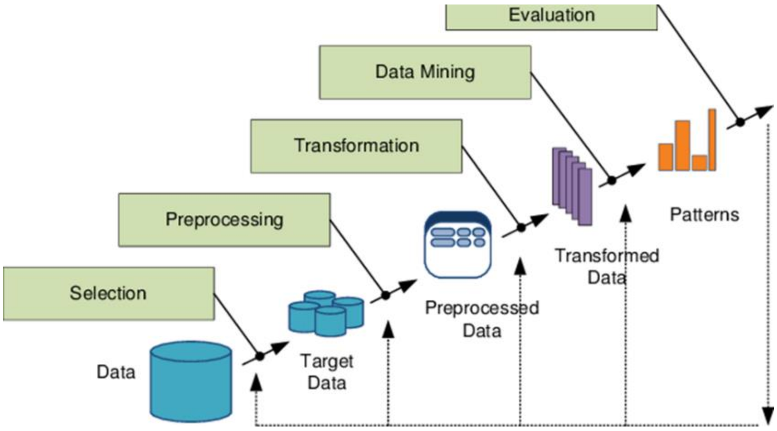


Figure 1. KDD Stage of Research [11]

Selection

At this stage, a data selection was performed on the data of currently active students and students who graduated with two variables, namely the predictor variable and the target variable. The target variables are the classification of students who graduated timely, in 4 years or less, with a minimum GPA of 3.0. The predictor variables are gender and GPA in the 3rd, 4th, 5th, and sixth semesters.

Preprocessing

All graduates from all departments in Al-Hidayah University Bogor were used as data. A cleaning data was performed on that data to examine the missing value, data duplication, or outlier data.

Transformation

After data cleaning, the next stage was data transformation based on the data type, where the data was a classification based on the category. The following Table 1 shows the categorization for predictor and target variables.

Table 1. Predictor Variable and Target Variable Categorization

Attribute	Information
Gender	Male/Female
GPA	0.00 – 4.00
Job Status	Employed/Unemployed
Graduation Prediction	Timely/Not

Data Mining

At this stage, the proper data mining technique was selected. For the classification function, the C4.5 and Support Vector Machine (SVM) was utilized. Classification is supervised learning, so this stage is in the supervised learning model [12].

Evaluation

This stage was performed to evaluate the prediction results of the Algorithm that have relative values with the actual data classification. The Confusion Matrix method was used as the evaluation method. The performance assessment values were accuracy dan error.

3. RESULTS AND DISCUSSION

3.1. *Data Selection*

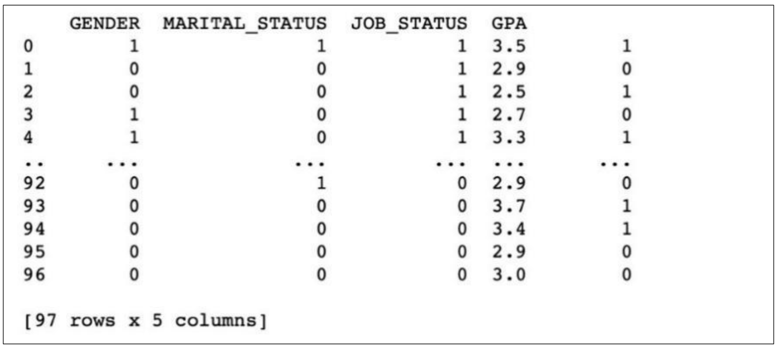
The source of raw data in this study is students from all departments from 2016 to 2019. The data source of the GPA is the student transcripts available in the academic database. The graduation data was taken from the data of students who graduated every year. The amount of data used in this study are 97 data from all departments and graduates from every department.

3.2. Data Preprocessing and Transformation

The data preprocessing was performed as follows: eliminate the lost data by eliminating incomplete data and predicting the lost data based on the suitable value. In this case, there was one blank gender data indicated as a male by examining the student's name.

After preprocessing, data transformation was performed. The male was transformed into 1, and the female was converted into 0. For marital status, married was transformed into one, and not married was transformed into 0. For job status, employed was transformed into one and unemployed was transformed into 0. The timely graduation prediction was converted into one, and not timely was transformed into 0. The transformation was not performed on GPA.

After performing preprocessing data and transformation, there are a total of 97 data with the following factors: gender, marital status, job status, GPA, and prediction, as shown in Figure 2:



```
    GENDER  MARITAL_STATUS  JOB_STATUS  GPA     
0        1               1           1  3.5    1
1        0               0           1  2.9    0
2        0               0           1  2.5    1
3        1               0           1  2.7    0
4        1               0           1  3.3    1
..     ...             ...         ...  ...  ...
92       0               1           0  2.9    0
93       0               0           0  3.7    1
94       0               0           0  3.4    1
95       0               0           0  2.9    0
96       0               0           0  3.0    0

[97 rows x 5 columns]
```

Figure 2. Research Data

Note:
Gender = 1 male, 0 female
Marital_status = 1 married, 0 not married
Job_status = 1 employed, 0 unemployed
GPA = IPK
Prediction = 1 timely, 0 not timely

3.3. Data Mining

In **the SVM Algorithm, a data conversion was done from nominal into numeric because the SVM Algorithm in Rapid Miner software can only be processed** using numeric data. After selecting the feature subset, a validation test was performed with 10-cross-validation done using the SVM model. The conversion process from nominal to numerical and cross-validation can be seen in Figure 3:

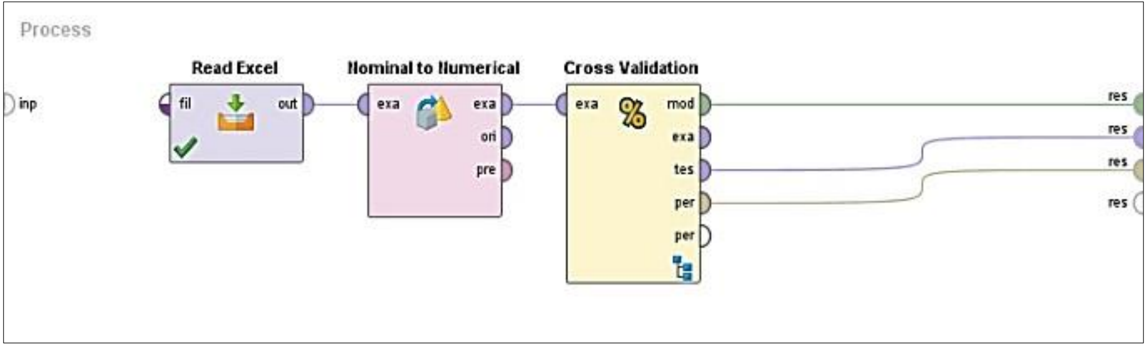
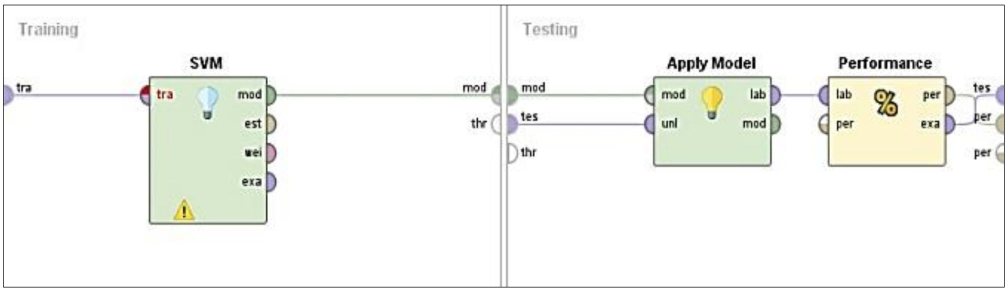


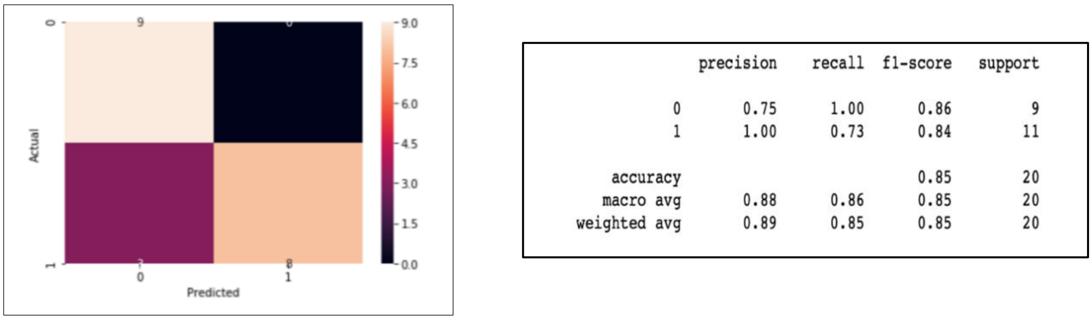
Figure 3. Data Mining Process Using SVM

The data training using the SVM Algorithm in the classification technique using Rapid Miner can be seen in Figure 4.



Gambar 4. Classification Model Using SVM Algorithm

In evaluating SVM Algorithm performance, Confusion Matrix reference was used. Confusion Matrix represented the prediction and the actual condition of the data. Based on the Confusion Matrix, the Accuracy, Precision, Recall, and Specificity can be determined. The results of the SVM Algorithm are confusion matrix, precision, recall, fi-score, support, and accuracy, as shown in Figure 5:



```
              precision    recall  f1-score   support

           0       0.75      1.00      0.86         9
           1       1.00      0.73      0.84        11

    accuracy                           0.85        20
   macro avg       0.88      0.86      0.85        20
weighted avg       0.89      0.85      0.85        20
```

Figure 5. SVM *Confusion Matrix, Accuracy, and Precision*

As shown in Figure 5, the precision and accuracy value was obtained. The precision rate of the attribute and its variables is 0.88 or 88%, and the accuracy rate of the attribute and its variables is 0.85 or 85%.

C4.5 Algorithm in the classification technique using Rapid Miner can be seen in Figure 6 as follows:

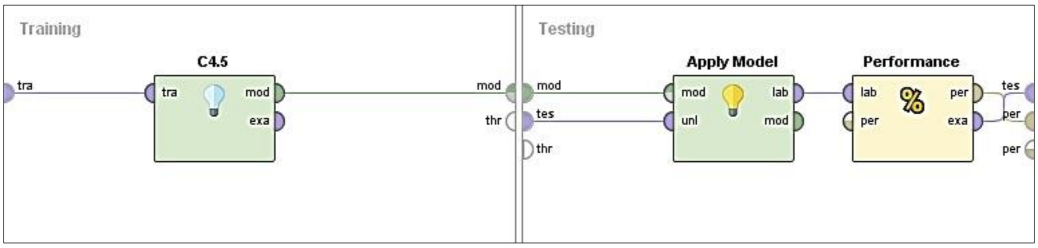
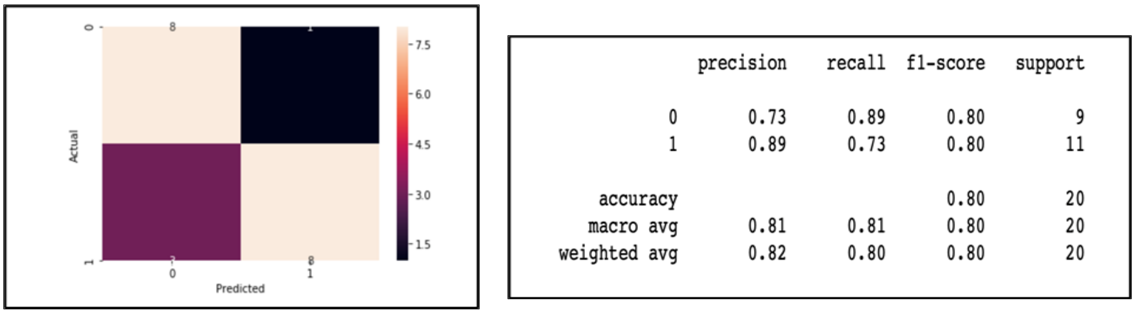


Figure 6. Classification Model Using C4.5 Algorithm

With the formula shown in Figure 6, the data obtained can be imported and processed using the C4.5 Algorithm for a decision-making process. The results of the C4.5 Algorithm are confusion matrix, precision, recall, fi-score, support, and accuracy table as figure 7:



```
              precision    recall  f1-score   support

           0       0.73      0.89      0.80         9
           1       0.89      0.73      0.80        11

    accuracy                           0.80        20
   macro avg       0.81      0.81      0.80        20
weighted avg       0.82      0.80      0.80        20
```

Figure 7. C4.5 Confusion Matrix, Accuracy, and Precision

As shown in the table, the precision rate of the attribute and its variables is 0,81 or 81%, and the accuracy rate of the attribute and its variables is 0,80 or 80%.

The visualization of the C4.5 decision tree was obtained based on entropy from each variable and Gain. The result is as Figure 8:

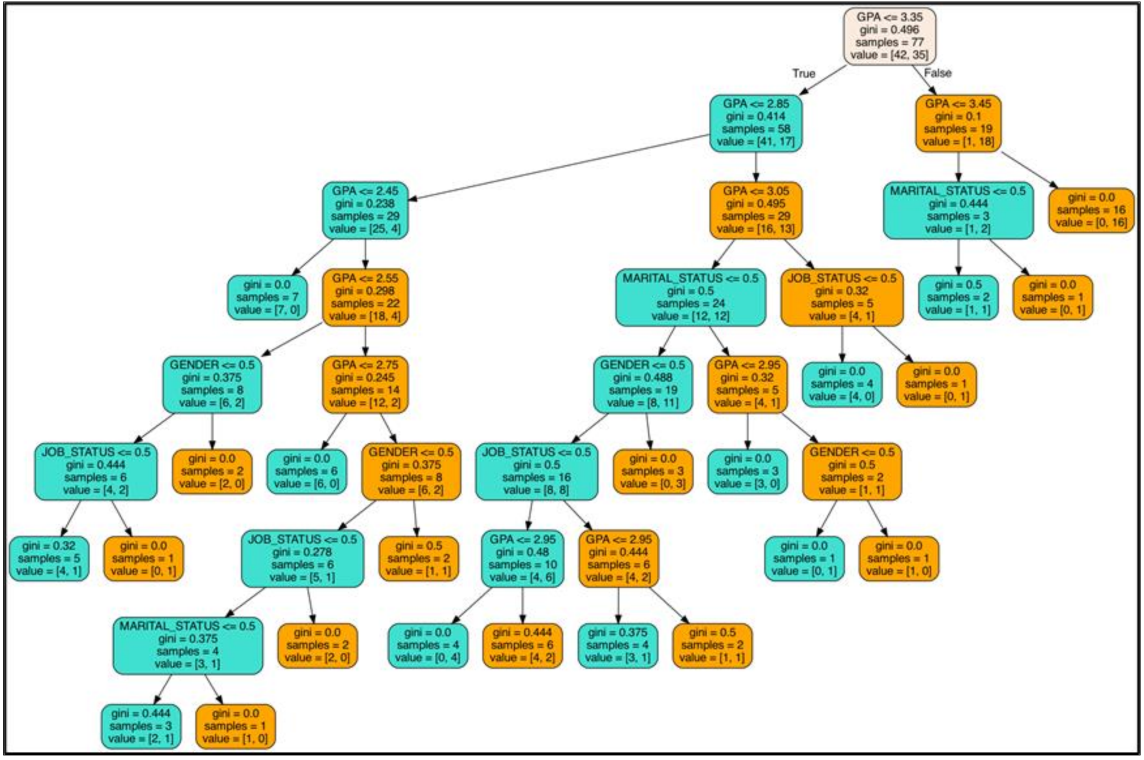


Figure 8. C4.5 Decision Tree Rule Visualization

The visualization in Figure 8 can be used as a model for new data that will be predicted. It gives a good and easy way to input new data as one of the classification rules.

3.4. Evaluation

Based on the calculation result of the Algorithm, the precision and accuracy can be compared as follows:

	Precision	Accuracy
SVM	0.88	0.85
C4.5	0.81	0.80

Table 2. Output Statistics of Precision and Accuracy Comparison of SVM and C4.5

The comparison table between both techniques are visualized as a histogram, as shown in Figure 9:

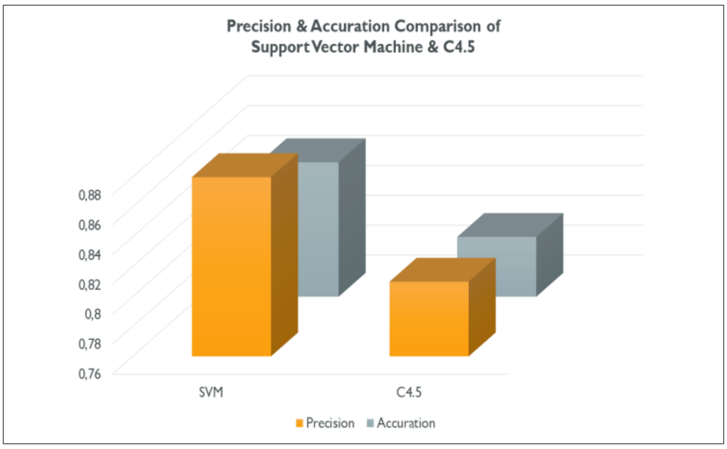


Figure 9. Precision & Accurate Comparison of Support Vector Machine (SVM) & C4.5

The comparison of accuracy and precision in 20% of data testing shows that SVM has higher accuracy of 85%, while C4.5 has 80% accuracy.

4. CONCLUSION

Based on the result of this study, it can be concluded that SVM and C4.5 data mining techniques can be used to predict the student graduation timeliness in Al-Hidayah University Bogor with gender, marital status, job status, and GPA as predictor factors because each technique generates good accuracy and precision (higher than 70%). The highest accuracy and precision rate are generated using SVM. However, C4.5 visualization can provide a good guide for rules in prediction. For future study, it is recommended to combine other Algorithm methods with the same or different attributes.

5. REFERENCES

[1] Suhardjono, G. Wiajaya, and H. Abdul, "Prediksi Waktu Kelulusan Mahasiswa Menggunakan Svm Berbasis Pso," *Bianglala Inform.*, vol. 7, no. 2, pp. 97–101, 2019.

[2] R. P. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, p. 1, 2018.

[3] A. Pratama, R. C. Wihandika, and D. E. Ratnawati, "Implementasi Algoritma Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. March, pp. 1704–1708, 2018.

[4] E. Supriyadi and D. I. Sensuse, "Optimasi Algoritma Support Vector Machine Dengan Particle Swarm Optimization Dalam Mendeteksi Ketepatan Waktu Kelulusan Mahasiswa : Studi Kasus Poltek Lp3i Jakarta 'Kampus Depok,'" vol. 1, no. 1, pp. 163–174, 2015.

[5] I. P. Astuti, "Prediksi Ketepatan Waktu Kelulusan Dengan Algoritma Data Mining C4.5," *Fountain Informatics J.*, vol. 2, no. 2, pp. 41–45, 2017.

[6] A. Rohman and M. Rochcham, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Mahasiswa," *Neo Tek.*, vol. 5, no. 1, 2019.

[7] C. N. Dengen, Kusrini, and E. T. Luthfi, "Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *J. Ilm. SISFOTENIKA*, vol. 10, no. 1, pp. 1–11, 2020.

[8] A. Fahrudin, L. Listiyoko, P. Surya, and A. Maksum, "Prediksi Peringkat Kelulusan Mahasiswa Untuk Menentukan Strategi Pemasaran Kampus Menggunakan Pohon Keputusan," no. November, 2017.

[9] I. Iskandar *et al.*, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4 . 5 Dengan Teknik Pruning," *J. Ilmu Komput. dan Sist. Inf.*, vol. 6, no. 1, pp. 64–68, 2018.

[10] A. Andie, "Penerapan Decision Tree Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Baru," *Technologia*, vol. 7, no. 1, pp. 8–14, 2016.

[11] Y. B. Samponu and K. Kusrini, "Optimasi Algoritma Naive Bayes Menggunakan Metode Cross Validation Untuk Meningkatkan Akurasi Prediksi Tingkat Kelulusan Tepat Waktu," *J. ELTIKOM*, vol. 1, no. 2, pp. 56–63, 2018.

[12] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed. Canada: Wiley-Interscience, 2014.