Predicting the Academic Performance of Students at Tay Bac University Using Decision Trees

Van-Tu Nguyen[1*], Thi-Quyen Tran[2]
[1]Faculty of Natural Sciences and Technology, Tay Bac University, Son La, Vietnam
[2] Faculty of Economic Engineering Son La College, Son La, Vietnam
*Corresponding author email: tuspttb@utb.edu.vn

Abstract:

Based on the admission information and first–year academic results collected from full–time students at Tay Bac University, this study applies classification techniques in data mining to build a prediction model for students' academic performance. The core of the study is the decision tree model, which enables analyzing influential factors and visualizing the prediction process in an intuitive manner. After evaluation and comparison, the model with the highest accuracy was selected and used to develop a prediction support tool, contributing to academic advising and helping students achieve better performance.

Keywords—**Decision Tree; Student Performance Prediction; Educational Data Mining; Classification Model**

I. INTRODUCTION

In higher education institutions, accurately predicting students' academic performance provides several advantages, such as identifying underperforming students for timely support or detecting outstanding students for scholarship consideration [1], [2]. With the development of technology and the increasing amount of educational data, the application of data mining techniques has become an effective approach in analysis and prediction [3].

In this study, the author applies the decision tree technique to predict the first–year academic performance of Tay Bac University students based on admission information, including gender, region, priority category, exam cluster, ethnicity, and total entrance exam score. The decision tree algorithm was chosen due to its intuitiveness, interpretability, and efficiency when dealing with classification–type data [2], [4].

II. RESEARCH CONTENT

A. *Analysis and model development*

The data mining process is implemented according to the standard model commonly used in data mining studies [1], [2], including data collection, cleaning, filtering, transformation into a suitable format, applying mining techniques to build the model, and evaluating the results obtained. The experimental data for prediction were collected from various sources. The annual admissions data include: candidate ID, gender, region, priority category, ethnicity, exam cluster, major, scores of each subject, priority points, and admission result (accepted or rejected). Another important source is the students' semester academic records, including student ID, major, subjects taken, and academic results. As the data were collected from management units, their authenticity and accuracy are highly reliable.

As a result, the author collected 5,187 records containing course grades and personal/admission information for 5,187 full–time undergraduate students at Tay Bac University. These Excel files were imported into the SQL Server 2008 database using the Import function. Then, they were organized into a usable structure for data mining by designing and executing SQL queries. The tool used for data mining was Microsoft Business Intelligence Development Studio 2008.

B. *The student performance prediction problem*

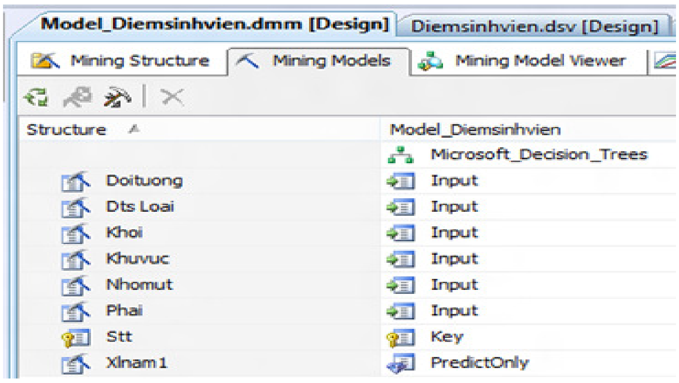
The prediction task aims to classify students into groups representing their first–year academic performance based on input attributes. These attributes include: (1) Admission information: gender, exam cluster, region, priority category, ethnicity, total entrance exam score. (2)First–year academic performance: classification levels: Excellent, Very Good, Good, Fairly Good, Average, Weak, Poor. After organizing the data into a suitable format, the next step is to study and select a model to classify students into the corresponding categories. This type of classification model aligns with the characteristics of classification problems in data mining [1], [4].

C. *Model construction and selection*

Among the common classification techniques, in addition to traditional methods such as Neural Networks, Logistic Regression, and Naive Bayes, several studies also mention classification based on association rule mining to enhance model interpretability [6]. However, given the nature of the dataset and the prediction objective, the decision tree algorithm was selected due to its intuitive structure, high interpretability, and suitability for categorical data [2], [4]. The dataset was divided based on the standard ratio: 70% for model training, 30% for testing and evaluation. The input attributes include gender, ethnicity, exam cluster, region, priority category, and total entrance exam score.

D. *Building the data mining model*

The data mining model named Model_DiemSinhVien was created in Microsoft Business Intelligence Development Studio following the guidelines from the technical reference in [1], [5]. The Microsoft_Decision_Trees algorithm was used. The default parameters ensure appropriate splitting levels, avoiding overfitting.



Structure	Model_Diemsinhvien
	Microsoft_Decision_Trees
Doituong	Input
Dts Loai	Input
Khoi	Input
Khuvuc	Input
Nhomut	Input
Phai	Input
Stt	Key
Xlnam1	PredictOnly

Fig. 1. *Data mining model illustration*

E. *Training the data mining model*

After the model was created, the training dataset was used to construct the decision tree. The algorithm divides the data based on input attributes to maximize information gain, similar to the machine learning processes described in [1], [2]. This process is performed automatically and iteratively until the model converges.

F. *Browsing the content of the data mining model*

Once the data mining model has been built and trained, its content can be examined in Analysis Manager. The model content represents the patterns discovered by the data mining algorithm in the dataset.

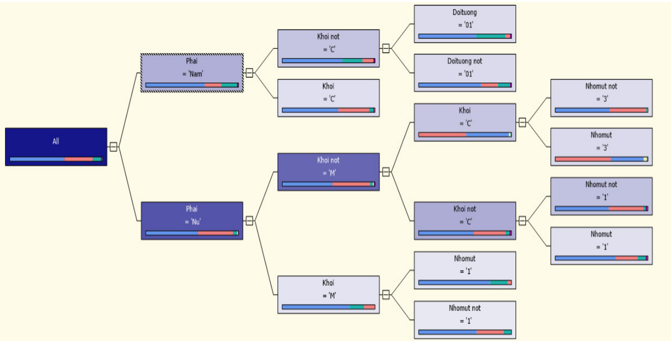


Fig. 2. *Decision tree used for prediction*

By browsing the model content, several rules can be extracted.

Rule 1: If Gender = “Male” and Exam Cluster ≠ “C” and Priority Category = “01”, then:

TABLE I. DISTRIBUTION OF ACADEMIC PERFORMANCE ACCORDING TO RULE 1 OF THE DECISION TREE

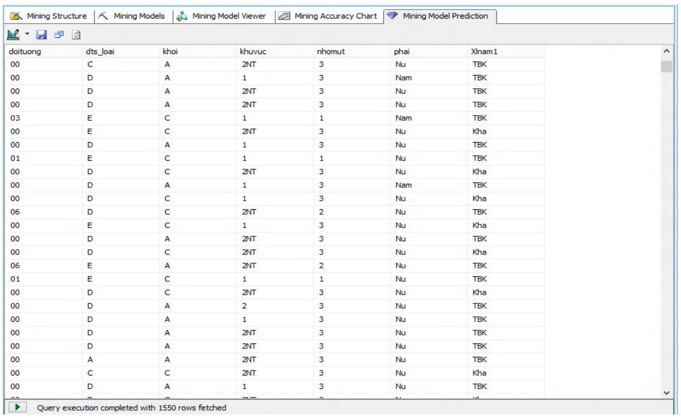
Classification	Cases	Percentage (%)
Very Good	0	0
Good	22	5.25
Fairly Good	257	60.98
Average	134	31.81
Weak	8	1.93

These rules are similar to those found in Decision Tree–based models described in [1], [4].

G. *Using the data mining model for prediction*

Prediction is an important task in data mining. It requires two components: a trained data mining model and a set of new cases. The prediction results consist of a new set of records containing values for predicted attributes and remaining input columns.

ISSN: **3107-6513**



doituong	dts_loai	khoi	khuvuc	nhomut	phai	Xlnam1
00	C	A	2NT	3	Nu	TBK
00	D	A	1	3	Nam	TBK
00	D	A	2NT	3	Nu	TBK
00	D	A	2NT	3	Nu	TBK
03	E	C	1	1	Nam	TBK
00	E	C	2NT	3	Nu	Kha
00	D	A	1	3	Nu	TBK
01	E	C	1	1	Nu	TBK
00	D	C	2NT	3	Nu	Kha
00	D	A	1	3	Nam	TBK
00	D	C	1	3	Nu	Kha
06	D	C	2NT	2	Nu	TBK
00	E	C	1	3	Nu	Kha
00	D	A	2NT	3	Nu	TBK
00	D	C	2NT	3	Nu	Kha
06	E	A	2NT	2	Nu	TBK
01	E	C	1	1	Nu	TBK
00	D	C	2NT	3	Nu	Kha
00	D	A	2	3	Nu	TBK
00	D	A	1	3	Nu	TBK
00	D	A	2NT	3	Nu	TBK
00	D	A	2NT	3	Nu	TBK
00	A	A	2NT	3	Nu	TBK
00	C	C	2NT	3	Nu	Kha
00	D	A	1	3	Nu	TBK

Query execution completed with 1550 rows fetched

Fig. 3. *Prediction results using Model_DiemSinhVien*

The results in Figure 3 show that, based on input information (priority category, type of score, exam cluster, region, priority group, gender) and the Model_DiemSinhVien, the corresponding values for the *xlnam1* attribute were obtained.

H. Evaluating the model's accuracy

The accuracy of the prediction model was evaluated to determine how well it performs. The Mining Accuracy Chart function was used for this purpose. The testing dataset—separated from the original dataset and not used in model training—was used for evaluation. Microsoft Business Intelligence Development Studio also generated an ideal model (Ideal Model) for comparison. The two main tools for evaluating model accuracy were the Lift Chart and the Classification Matrix. In this study, the model achieved an accuracy of over 76%.

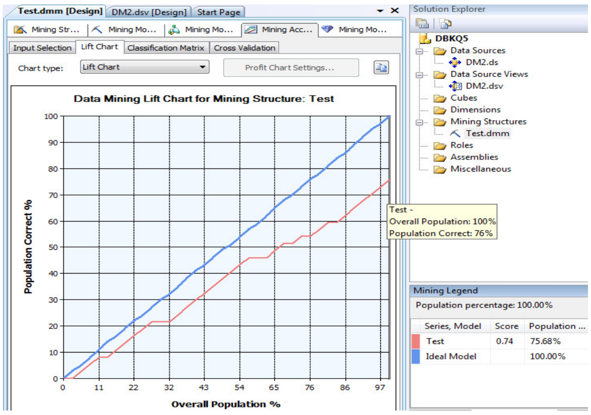


Chart type: Lift Chart

Test -
Overall Population: 100%
Population Correct: 76%

Population percentage: 100.00%

	Series, Model	Score	Population ...
🟥	Test	0.74	75.68%
🟦	Ideal Model		100.00%

Fig. 4. *Accuracy of the prediction model*

III. CONCLUSION

Based on the initial training dataset, the data mining model built in this study enables analysis of factors influencing students' academic performance, the degree of impact of each input attribute, and prediction of academic outcomes based on given input information. To apply the prediction model more effectively in practice, it is necessary to continue expanding the dataset, implement prediction regularly, conduct real–world validation, and evaluate the results frequently. This approach is considered practical and appropriate. Future research directions may include applying additional models to the student performance prediction problem, such as predicting second–year performance using first–year information or evaluating overall academic performance using multi–year data.

REFERENCES

[1]J. MacLennan, Z. Tang, and B. Crivat, Data Mining with Microsoft SQL Server 2008, Wiley Publishing, 2008.

[2]J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Elsevier, 2006.

[3]B. Knight, D. Knight, A. Jorgensen, P. LeBlanc, and M. Davis, Knight's Microsoft Business Intelligence 24-Hour Trainer, Wiley Publishing, 2010.

[4]"Khai phá dữ liệu (Data Mining)," bis.net.vn. Available: http://bis.net.vn/forums/p/366/628.aspx#628.

[5]"Developing Application that uses Analysis Services," Microsoft Developer Network. Available:http://social.msdn.microsoft.com/Forums/zh/sqldatamining/thread/fb74ab56-1172-4460-8953-f566ca0a0cf3.

[6]M. Nofal and S. Bani-Ahmad, "Classification based on association rules mining techniques: a general survey and empirical comparative evaluation," UBICC Journal. Available: http://www.ubicc.org/files/pdf/507_507.pdf