Predicting the Professional Field of Students Using Data Mining: A Case Study of an Autonomous University in Thailand

Phichayasini Kitwatthanathawon

The Institute of Digital Arts and Science, Suranaree University of Technology, Thailand
Email: pichak@sut.ac.th (P.K.)
Manuscript received January 29, 2024; revised May 14, 2024; accepted July 8, 2024; published September 14, 2024

***Abstract*—At the Institute of Digital Arts and Sciences at Suranaree University of Technology, students face the crucial decision of choosing between digital technology and digital communication as their professional field. This choice significantly influences their academic pursuits and future career paths. This research aimed to construct and compare the performances of various models in predicting students' selection of professional fields, utilizing data from student questionnaires at the Institute of Digital Arts and Science. Classification techniques, considered a subset of data mining methods, were applied, and models were constructed using five algorithms: Decision Tree, Naïve Bayes, One Rule (OneR), Support Vector Machine, and K-Nearest Neighbors. These models were evaluated based on accuracy, recall, precision, and F-measure and cross-validated with 10, 20, and 30-fold evaluations. The findings revealed that the Naïve Bayes algorithm-based model, especially with 20-fold cross-validation, was most accurate, achieving 89.6%. The Support Vector Machine algorithm-based model exhibited the highest precision at 82.1% with 30-fold cross-validation. The Decision Tree algorithm-based model achieved the highest recall and F-measure at 83.3% and 81.5%, respectively, with 10-fold cross-validation.**

***Keywords*—professional fields, major selection, prediction, educational data mining, classification**

I. Introduction

From the past to the present, education has been considered the most fundamental basis, as it is the foundation for development in every aspect. In addition to acquiring knowledge and personal development, it also determines the direction of a country, especially higher education, as it aims to develop students to become leaders capable of effectively accommodating economic and social changes. Countries that promote education in the right direction, as well as have quality planning, inevitably possess efficient human resources. This results in that country having human resources as an asset with the potential to drive the country forward. Meanwhile, choosing a field of study is also crucial. Opting to study in a field that one likes or excels in often significantly impacts future career prospects [1].

The Institute of Digital Arts and Science (DIGITECH) at Suranaree University of Technology is an academic faculty providing digital technology education. It offers courses in Digital Technology and Digital Communication, which align with the current needs of the business sector and industry in accordance with the government's Thailand 4.0 policy and significant global trends [2].

The selection of professional fields is crucial for students as it directly affects their future career paths and may impact their lives. Students may lack experience, are not fully aware of their exact needs, and may not have insufficient knowledge about each professional field. They may choose based on personal preferences, follow their friends' or parents' opinions, or face issues related to family circumstances. This may lead to wasting time and opportunities when they realize they are not suited for the chosen field of study [3]. For these reasons, the researcher recognized the importance of developing a model to help predict undergraduate students' selection of professional fields of study. This research applied classification techniques, considered key data mining methods, to construct predictive models using the Decision Tree, Naïve Bayes, One Rule (OneR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms. These algorithms help analyze relevant factors and predict students' future professional fields. The objective of this research is to construct and compare the performances of the models in predicting professional fields.

II. Literature Review

A. Data Mining

Data mining is the process of discovering hidden patterns, approaches, and relationships in large datasets using machine learning, statistics, database systems, and pattern recognition. Data mining is a method of Knowledge Discovery in Databases (KDD), or in other words, a process that deals with data by analyzing existing data and extracting knowledge or key elements for use in analysis or various predictions [4]. There are various algorithms for data analysis, examples of which are as follows.

1) Naïve bayes

Naïve Bayes, or Bayes' Theorem, was developed by Thomas Bayes. It utilizes the principle of probability to develop the theorem [5] by addressing problems that arise to create conditions for new data classification with a straightforward algorithm. The Naïve Bayes principle uses probability calculations for outcome prediction, serving as one method for solving classification problems capable of predicting results. It analyzes the relationships between variables to construct conditional probability for each relationship. This algorithm is suitable for cases involving a large number of sample datasets where the characteristics of the samples are independent of each other.

2) Decision tree

A decision tree uses data to construct a predictive model in

the form of a tree [6]. It consists of 1) nodes, each representing a decision based on various characteristics. 2) Branches, which are values or results obtained from testing, serving as connections between nodes and illustrate all possible characteristics of each node. 3) Leaf nodes, which are the nodes at the lowest level of a decision tree. They represent a group of data or results obtained from the decision-making conditions. The decision tree is widely used because it is easy to understand and interpret, and it can manage both nominal and numerical data.

3) K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) is a machine learning algorithm used for both classification and regression tasks. It classifies data based on the nearest neighboring data. The algorithm works by finding the K closest data points to a given data point in the training data, using a distance function such as Euclidean, Manhattan, or Minkowski Distance. It then calculates the average or votes for predicting or classifying new data [7]. If the data of interest is closest to certain data, the system will provide an answer similar to the nearest data. It does not use the training data to construct a model but instead uses the nearest data to construct the model directly. The K-Nearest Neighbors algorithm is straightforward to understand and use but may face challenges in classifying complex data.

4) Support Vector Machines (SVM)

Support Vector Machine is a supervised learning algorithm focusing on finding the most realistic or probable vector to solve data classification problems [8]. It operates on the principle of determining the coefficients of an equation to create a line that best separates the groups of data input into the training process. Support Vector Machine excels at classifying complex and non-linear data by identifying the most feasible vector that is capable of accurately separating data groups with a maximum margin.

5) One Rule (OneR)

One Rule (OneR) is an algorithm that is easy to understand and use for rule creation. It operates on the principle of constructing classification rules by selecting a single attribute with the least error to predict the class of data. This approach results in a minimal number of classification rules [9].

B. Data Analysis Process with CRISP-DM

This is a standard process used in data mining, designed for analysis and business application [10]. The process is shown in Fig. 1.

From Fig. 1, the CRISP-DM data analysis process consists of 6 steps, which are:

Step 1: Business Understanding: This step involves comprehending the targeted business or organizational objectives by examining various business factors, and then converting these into a format that can be analyzed and used for operational planning.

Step 2: Data Understanding: This step starts with examining the collected data, followed by understanding it to select the appropriate data for analysis.

Step 3: Data Preparation: This step includes converting the data into a format ready for model construction.

Step 4: Modeling: This step involves constructing models using various algorithms to identify a model capable of solving the problem.

Step 5: Model Evaluation: This involves assessing the models' performances to ensure they are ready for deployment.

Step 6: Deployment: This step involves using the models that have produced the best results and passed performance evaluation for real-life applications to analyze and solve problems.

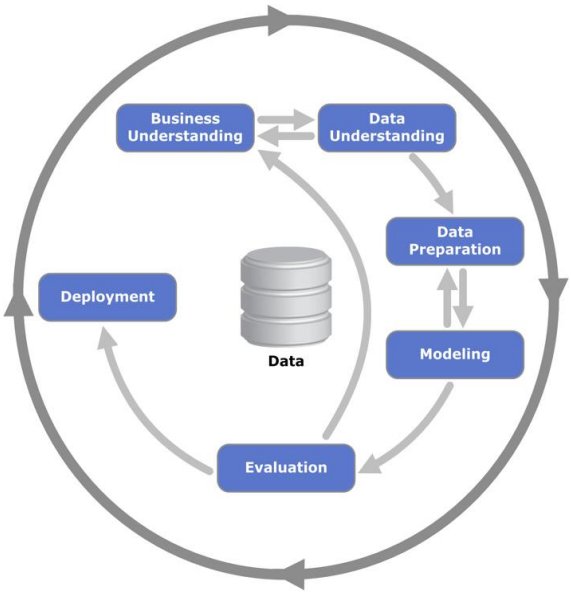

Fig. 1. Data analysis process with CRISP-DM.

C. Related Work

Existing research related to Educational Data Mining focuses on various areas including predicting student performance, personalizing learning experiences, identifying at-risk students, enhancing curriculum design, and analyzing student behavior. The details are as follows.

1) Predicting student performance

The research in this area often utilizes algorithms such as decision trees, neural networks, and Bayesian networks to forecast students' future performance based on historical data. These predictions help educators intervene early with students who may be struggling [11–15].

2) Identifying at-risk students

The research in this area aims to identify students who are at risk of dropping out by analyzing patterns in attendance, grades, and engagement. The research outcomes, early identification, allows institutions to provide targeted support to these students [16, 17].

3) Curriculum design

The research in this area aims to analyzes student performance data to identify which areas of the curriculum need improvement. This can lead to more effective instructional strategies and materials [18, 19].

4) Behavioral analytics

The research in this area studies student behavior patterns, including interaction with learning management systems and participation in online forums, to understand how these behaviors correlate with academic success [3, 20, 21].

5) Personalized learning

The research in this area aims to tailor learning experiences to individual students' needs. Techniques such as clustering and recommendation systems are used to identify the most effective learning paths and resources for each student [22–26].

Nakhipova *et al*. [13] developed and implemented the Naive Bayes methodology for predicting student performance based on their characteristics. The findings affirm that employing the Naive Bayes classifier in this context can be both effective and practical with an Accuracy score of 85%.

Al-Barrak *et al*. [14] predicted students' final Grade Point Average (GPA) based on their grades in previous courses. They collected students' transcript data that included their final GPA and their grades in all courses. After pre-processing the data, we applied the J48 decision tree algorithm to discover classification rules. They extracted useful knowledge for final GPA, and identify the most important courses in the students' study plan based on their grades in the mandatory courses.

Shayan and Zaanen [15] analyzed the prediction models of student performance from their online behavior based on Moodle Learning Management System (LMS) data. They employed the decision tree J48 and ID3 classification algorithms in order to enhance learners' achievement. The research results showed that prior GPA, midterm grade, the number of views, clicks, and sessions had an impressive impact on students' performance and lecturers have to pay more attention to this.

Peterson *et al*. [24] predicted students' future professional fields based on their academic performance and extracurricular activities using machine learning algorithms. They employed various machine learning models including Decision Trees, Random Forest, and Support Vector Machines (SVM) on a dataset of high school students' grades and activity logs. The research results showed that the Random Forest model achieved the highest accuracy of 85%, suggesting that academic performance and extracurricular activities are strong predictors of future professional fields.

Zhang *et al*. [25] predicted career trajectories of university graduates using machine learning techniques based on their academic records and internships. They utilized a combination of Neural Networks and K-Nearest Neighbors (KNN) on a dataset comprising university students' grades, internship experiences, and job placement records. The research results showed that the Neural Networks provided a prediction accuracy of 78%, highlighting the importance of internships in shaping career paths. However, this research did not integrate psychological assessments or mentorship influences, which could offer a more holistic view of career prediction.

Singh *et al*. [26] developed a machine learning model for predicting the professional fields of engineering students based on their coursework and project involvement. They implemented Logistic Regression, Naive Bayes, and Gradient Boosting on a dataset of engineering students' grades, project participation, and specialization preferences. The research results showed that the Gradient Boosting model outperformed others with an accuracy of 82%, indicating the significance of project-based learning in career predictions. The study lacked real-time labor market data integration, which could enhance the model's relevance and accuracy in predicting current job market trends.

In Thailand, Thepprasit and Sanrach [3] analyzed the factors affecting the selection of 15 majors in the Faculty of Education at Chiang Rai Rajabhat University. They collected data on 9 variables from undergraduate students through academic support and registration during the academic years 2013–2017, totaling 3,867 students. They analyzed the correlation of the variables using the data mining process with the Decision Tree classification technique. The research findings indicated that factors influencing major selection were the pre-enrollment study plan and gender. The accuracy measurement was as high as 72.5%, signifying that it is an efficient and reliable model.

Jaruteerapan [27] studied the factors influencing the choice of field of study at the bachelor's degree level in the Faculty of Management Science at Loei Rajabhat University. The study utilized data from 678 first-year students, gathered through a questionnaire, on factors influencing their choice of study in the Faculty of Management Science. Statistics used for data analysis included percentages, averages, standard deviations, and Analysis of Variance (ANOVA). The research found that, on average, factors influencing the choice of field of study in the Faculty of Management Science at Loei Rajabhat University were considered significant. The study found that students considered all 5 factors, namely the characteristics of the field of study, publicity and public relations, location, influence on decision-making, and the cost associated with the chosen field of study, important in their choice of field of study. These factors were statistically significant at the 0.05 level.

Wongthong *et al*. [28] constructed and compared the performance of the data correlation models using the Artificial Neural Network, Na ïve Bayes, and Decision Tree algorithms by collecting data from 395 computer professionals. The data was divided into 5 parts and analyzed by measuring its accuracy and precision. It was found that the most efficient model was the model obtained from the Na ïve Bayes algorithm, which had an accuracy value of 79.64% and a precision value of 74.04%, followed by the Artificial Neural Network algorithm, which had an accuracy value of 77.82% and a precision value of 73.60%. The Decision Tree algorithm had an accuracy value of 54.69% and a precision value of 38.72%, respectively.

Panyawong [29] studied the factors influencing the choice of major to study at the bachelor's degree level and constructed a model for predicting graduation by collecting student data from the Faculty of Architecture in the academic year 2014–2016, totaling 929 samples. Using the OneR algorithm to create a predictive model, the performance of the model was measured in terms of accuracy and recall. The model had similar high accuracy and recall values at 91% and 91.1%, respectively.

Wannaprapha [30] conducted a study to construct a model for screening individuals likely to succeed in undergraduate studies through data mining. He collected graduation data of 738 graduates from the registration and education processing system, which was classified into 30 attributes. The study

utilized the Decision Tree, Support Vector Machine, and Deep Learning algorithms. The dataset was divided for training and testing in ratios of 70:30, 75:25, and 80:20, respectively. The study's results revealed that the Deep Learning algorithm achieved the highest accuracy at 85.94% when the training and testing processes were divided in an 80:20 ratio.

Jareanying [31] did educational data mining by classifying data to find correlations among variables and comparing the performances of the Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (K-NN) algorithms using data from 649 secondary school students, divided into 31 attributes. The research results showed that the Decision Tree, Random Forest, and Naïve Bayes achieved an accuracy value of 91.54%, while the K-Nearest Neighbors (K-NN) algorithm had an accuracy value of only 82.31%.

Most existing research focuses on using machine learning to predict student performance using different algorithms, which can be summarized as shown in Table 1.

Table 1. A comparison of research related to educational data mining

Research	Research Area*	Algorithm/Technique										
		Statistics Analysis	Logistic Regression	Gradient Boosting	Decision Tree	Support Vector Machine	One Rule	Naïve Bayes	Artificial Neural Network	K-Nearest Neighbors	Random Forest	Deep Learning
[3]	B				✓							
[13]	P		✓	✓	✓			✓	✓		✓	
[14]	P				✓							
[15]	P				✓							
[24]	L				✓	✓					✓	
[25]	L								✓	✓		
[26]	L		✓	✓				✓				
[27]	B	✓										
[28]	B				✓			✓	✓			
[29]	B						✓					
[30]	P				✓	✓						✓
[31]	P				✓			✓		✓	✓	
This Research	L				✓	✓	✓	✓		✓		

* P = Predicting Student Performance, I = Identifying At-Risk Students, C = Enhancing Curriculum Design, B = Analyzing Student Behavior, L = Personalizing Learning Experiences

From studying related research, it can be concluded that most research on the educational data mining utilized the Decision Tree and Naïve Bayes algorithms, which provide high predictive performance. In Thailand, most existing research focuses on analyzing student behavior and predicting student performance. Limited research has incorporated psychological profiles or personality traits, which are crucial for a comprehensive career prediction model. Moreover, identifying the most relevant features for predicting professional fields is complex and often domain-specific. There is a need for more sophisticated methods to integrate diverse data sources, such as educational records, job histories, and personal interests. Therefore, integrating personal interests and conducting diverse experiments is a way to ensure that machine learning models are accurate, fair, and useful in real-world applications. This research utilized the aforementioned algorithms, as well as other widely used algorithms, namely OneR, K-Nearest Neighbors, and Support Vector Machine, to construct models and compare their performances in predicting the professional fields of students at the Institute of Digital Arts and Science. The objective was to find the most appropriate algorithm for classifying students' professional fields and to analyze factors influencing the selection of these professional fields. This model will serve as a tool to aid individuals interested in further education in choosing their field of study.

III. Research Methodology

To construct a model for predicting the selection of professional fields for students at the Institute of Digital Arts and Science, the data analysis process using CRISP-DM, a standard method for data mining, was applied to suit the context of this research as follows.

A. Problem Understanding

Understanding the problem regarding students' selection of professional fields was essential for investigating the issue, gathering relevant data, and defining the research objective in analyzing the factors influencing each student's selection of a professional field.

B. Data Understanding

Data understanding was a step involving collecting individual student data and personal interests using a questionnaire, which included 15 questions that had been verified and assessed for consistency (Index of Consistency: IOC) by 3 experts. The questions were categorized into personal factors, financial factors, and occupational factors, as shown in Table 2. A total of 310 respondents completed the questionnaire. The data was then used to construct models using Naïve Bayes, Decision Tree, K-Nearest Neighbors, and Support Vector Machine algorithms.

C. Data Preparation

This stage involved exploring and verifying the data collected from the previous step to prepare it for analysis and model construction. The details are as follows:

1) Adjusting the program to suit the dataset

For this research, data was analyzed using the Weka 3.8.5 program. However, since the imported data was in Thai, it was necessary to adjust the program in terms of file encoding to support the Thai language.

2) Discretization

Converting the data type of every attribute in occupational factors from numeric to nominal data.

3) Data cleansing

Verifying and correcting erroneous data by removing data

records with unclear or missing information, eliminating spaces that are mixed with data, and correcting spelling errors.

The data preparation process resulted in a dataset ready for model construction, consisting of 226 records with 15 attributes, as shown in Table 2.

Table 2. Details of the input data

No.	Question Categories	Description	Data Type	Remarks
1	Personal Factor	Gender	Nominal	Male, Female, Others
2	Financial Factor	Monthly family income	Numeric	
3	Occupational Factor 1	Ease of finding a job after completing studies in the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
4	Occupational Factor 2	Opportunity for high compensation in the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
5	Occupational Factor 3	Ability to earn extra income in various ways in the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
6	Occupational Factor 4	Likely to experience increasing job market demand in the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
7	Occupational Factor 5	High importance given by organizations to the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
8	Occupational Factor 6	Job positions available in public and private sectors	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
9	Occupational Factor 7	Alignment with current and future economic conditions	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
10	Occupational Factor 8	Applicability of field knowledge to nearly all professions	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
11	Occupational Factor 9	Potential for rapid advancement in the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
12	Occupational Factor 10	Ability to provide life stability through the field	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
13	Occupational Factor 11	Office-based work environment and atmosphere	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
14	Occupational Factor 12	Opportunities for constant intellectual involvement and thinking	Nominal	Very Interested, Somewhat Interested, Neutral, Not Very Interested, Not at All Interested
15	Professional Field Group	Professional fields	Nominal	DT – Digital Technology DC – Digital Communication

D. Modeling

The process involved using the training data to find patterns of relationships within the dataset using standard algorithms to achieve the best results in predicting students' professional fields, namely the Decision Tree, Naïve Bayes, OneR, K-Nearest Neighbors, and Support Vector Machine. This research utilized the Weka 3.8.5 program as a tool and employed algorithms such as J48, Naïve Bayes, OneR, Instance Based for K-Nearest Neighbor (IBK), and Sequential Minimal Optimization (SMO), respectively. The dataset for model construction, ready for analysis, consisted of 226 records divided into training data and test data. This research used 10-fold, 20-fold, and 30-fold cross-validation. In the first round, the first part of the data was used as test data, and the remaining as training data for model construction. This process was repeated, with each round alternating different parts as the test data until all parts had been used.

E. Model Evaluation

This research conducted cross-validation using the test data by dividing it into 10-fold, 20-fold, and 30-fold to calculate the performance of the models used in predicting students' selection of professional fields. The model that had the highest performance was selected. This performance was calculated based on the Confusion Matrix with the prediction of students' professional field selection and their actual chosen professional fields. The types of data used for evaluation can be categorized as shown in Table 3, namely:

1) Data where the model predicted that students would select the professional field of Digital Technology (DT), and the actual value was that the students selected that field (True Positive: TP).
2) Data where the model predicted that students would select the professional field of Digital Technology (DT), but the actual value was that the students did not select that field (False Positive: FP).
3) Data where the model predicted that students would not select the professional field of Digital Technology (DT), and the actual value was that the students did not select that field (True Negative: TN).
4) Data where the model predicted that students would not choose the professional field of Digital Technology (DT), but the actual value was that the students selected that field (False Negative: FN).

Table 3. Confusion matrix

		Predicted Value	
		Digital Technology (DT)	Digital Communication (DC)
Actual Value	Digital Technology (DT)	TP	FN
	Digital Communication (DC)	FP	TN

The values from the Confusion Matrix can be used to calculate the models' performances, namely accuracy, recall, precision, and F-measure, as shown in Eqs. (1)–(4) [32].

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100\% \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \times 100\% \quad (3)$$

$$\text{F-measure} = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

IV. Result

This research utilized a dataset from a questionnaire survey of students from the Institute of Digital Arts and Science, comprising 226 records with 15 attributes. Classification techniques in data mining were used to construct models using the Decision Tree, Naïve Bayes, OneR, Support Vector Machine, and K-Nearest Neighbors algorithms. The test data was divided into 10-fold, 20-fold, and 30-fold for cross-validation. The models' performances were evaluated based on precision, recall, accuracy, and F-measure. The results of the performance analysis of the models are presented in Tables 4 and 5.

Table 4. Precision and recall

Algorithm	Precision			Recall		
	10-fold	**20-fold**	**30-fold**	**10-fold**	**20-fold**	**30-fold**
Decision Tree	**79.8%**	75.0%	74.9%	**83.3%**	70.0%	80.8%
Naïve Bayes	80.0%	**81.0%**	80.2%	77.5%	**80.0%**	79.2%
OneR	79.3%	**79.6%**	79.4%	65.0%	66.7%	**67.5%**
Support Vector Machine	81.5%	81.4%	**82.1%**	**73.3%**	71.7%	**73.3%**
K-Nearest Neighbors	**69.1%**	69.0%	67.5%	64.2%	**67.1%**	60.0%

Table 5. Accuracy and F-Measure

Algorithm	Accuracy			F-Measure		
	10-fold	**20-fold**	**30-fold**	**10-fold**	**20-fold**	**30-fold**
Decision Tree	**89.3%**	84.8%	85.4%	**81.5%**	72.4%	77.7%
Naïve Bayes	88.7%	**89.6%**	89.0%	78.7%	**80.5%**	79.7%
OneR	86.7%	86.4%	**87.0%**	71.2%	72.3%	**72.8%**
Support Vector Machine	88.3%	88.7%	**89.3%**	77.1%	76.1%	**77.4%**
K-Nearest Neighbors	**80.6%**	**80.6%**	80.0%	**66.5%**	65.0%	63.4%

From Table 4, it can be seen that the model using the Support Vector Machine algorithm had the highest precision of 82.1% when the test data was divided into 30-fold for cross-validation. Next was the Naïve Bayes algorithm, with a precision of 81.0% when the test data was divided into 20-fold for cross-validation. The Decision Tree algorithm had a precision of 79.8% when the test data was divided into 10-fold for cross-validation. The OneR algorithm had a precision of 79.6% with 20-fold, and the K-Nearest Neighbors algorithm had the lowest precision of 69.1% with 10-fold for cross-validation.

Regarding recall, the model from the Decision Tree algorithm had the highest recall of 83.3% when the test data was divided into 10-fold for cross-validation. Next was the Naïve Bayes algorithm, with the recall of 80.0% when the test data was divided into 20-fold for cross-validation. The Support Vector Machine had a recall of 73.3% when the test data was divided into both 10 and 30-fold for cross-validation. The OneR algorithm had a recall of 67.5% with 30-fold for cross-validation, and the K-Nearest Neighbors algorithm had the lowest recall of 67.1% with 20-fold for cross-validation.

From Table 5, it can be seen that the model using the Naïve Bayes algorithm had the highest accuracy of 89.6% when the test data was divided into 20-fold for cross-validation. Next were the Decision Tree and Support Vector Machine algorithms, with equal accuracy of 89.3% when the test data was divided into 10 and 30-fold, respectively. The OneR algorithm had an accuracy of 87.0% when the test data was divided into 30-fold for cross-validation. The K-Nearest Neighbors algorithm had the lowest accuracy at 80.6% when the test data was divided into 10 and 20-fold for cross-validation.

Regarding the F-measure, it can be seen that the model from the Decision Tree algorithm had the highest F-measure of 81.5% when the test data was divided into 10-fold for cross-validation. Next was the Naïve Bayes algorithm, with an F-measure of 80.5% when the test data was divided into 20-fold for cross-validation. The Support Vector Machine had an F-measure of 77.4% when the test data was divided into 30-fold for cross-validation. The OneR algorithm had an F-measure of 72.8% with 30-fold for cross-validation, and the K-Nearest Neighbors algorithm had the lowest F-measure of 66.5% with 10-fold for cross-validation.

V. Discussion

When considering the best performance values of all algorithms, it can be concluded that the model using the Decision Tree algorithm shows good performance among other algorithms. The precision of 79.8% indicates its ability to predict professional field of students with high accuracy. The highest recall of 83.3% means that 83.3% of students selected the professional field of Digital Technology (DT) were correctly identified by the model. The accuracy of 89.3% means that 89.3% of the predictions of students selected the professional field of Digital Technology (DT) using this algorithm were correct. The 81.5% of F-measure indicates a good balance between precision and recall. This is

important for identifying students who need additional support. Thus, the model using the Decision Tree algorithm was chosen because the algorithm is able to accurately predict professional field of student and is good at identifying students selected the professional field of Digital Technology (DT). Also, the Decision Tree algorithm copes well with categorical data, has an understandable structure, can model complex relationships, and is effective on small dataset, which is suitable for this research.

Dividing the test data into 30-fold for cross-validation resulted in higher accuracy and precision values. Meanwhile, dividing the test data into 10 and 20-fold for cross-validation resulted in higher recall and F-measure values, respectively. Additionally, when considering the average performance of all algorithms, it was found that regardless of whether the test data was divided into 10, 20, or 30-fold for cross-validation, the performance values were very similar, with approximately 1% differences. However, dividing the test data into 10-fold for cross-validation resulted in higher performance compared to 20 and 30-fold.

In addition, when considering the rule derived from the Decision Tree model, which is easy to understand and provided the highest recall and F-measure values, the model demonstrated the relationship between questionnaire factors important to students during professional field selection. For instance, students in the field of Digital Technology showed interest in the current economic situation and were more likely to choose professions supported by both government and private sector positions compared to students in the field of Digital Communication. Furthermore, students in the field of Digital Technology were highly interested in choosing professional fields where jobs are readily available upon graduation and offer high compensation. Similarly, students in the field of Digital Communication were interested in choosing professional fields supported by both government and private sector positions, where jobs are readily available upon graduation and offer high compensation. However, students in the field of Digital Communication placed more importance on choosing professional fields where organizations give significant importance to the field. They also showed more interest in professions with working environments and atmospheres that are office-based compared to students in the field of Digital Technology. The aforementioned rule can assist in decision-making for students who have not yet chosen a professional field or are deciding on a professional field to study in the future.

VI. Conclusion

This research constructed and compared the performances of various models in predicting students' selection of professional fields at the Institute of Digital Arts and Science. It employed classification techniques, considered a subset of data mining methods, and constructed models using algorithms such as Decision Tree, Naïve Bayes, OneR, Support Vector Machine, and K-Nearest Neighbors. The test data was divided into 10, 20, and 30-fold for cross-validation. The models were evaluated based on precision, recall, accuracy, and F-measure. The performance analysis of the models revealed that the Naïve Bayes algorithm produced the model with the highest accuracy of 89.6% when the test data was divided into 20-fold for cross-validation. The Support Vector Machine algorithm produced the model with the highest precision of 82.1% when divided into 30-fold for cross-validation. The Decision Tree algorithm produced the model with the highest recall and F-measure of 83.3% and 81.5%, respectively, when the test data was divided into 10-fold for cross-validation. Hence, the Decision Tree algorithm model was selected because of its high accuracy in predicting the professional fields of students. The results obtained from the data analysis in this research are models with the highest performance for predicting the professional field selection of students at the Institute of Digital Arts and Science. These models can assist in decision-making for individuals deciding on selecting future professional fields and also serve as guidelines for those interested in data analysis using data mining techniques.

To make the model more usable, future research could involve increasing the number of attributes and the number of records in the dataset used for model construction. This could include tuning related parameters to enhance the model's accuracy. Additionally, employing a variety of modeling algorithms could lead to greater diversity and improved performance.

Conflict of Interest

The author declares no conflict of interest.

References

[1] J. Suknoi, "Decision-making to further study at a higher education level of high school students at Thasala Prasitsuksa School," M.S. thesis, Faculty of Business Administration Program in Finance and Banking, Ramkhamhaeng University, 2020.

[2] Institute of Digital Arts and Science, Suranaree University of Technology. *About the Institute of Digital Arts and Science*. [Online]. Available: https://digitech.sut.ac.th/about-program.php

[3] R. Thepprasit and C. Sanrach, "The analysis of factors affecting choosing a major of undergraduate students of the Faculty of Education by using data mining technique," *Journal of Graduate Studies Valaya Alongkorn Rajabhat University*, vol. 14, no. 1, pp. 134–146, 2020.

[4] S. Sinsomboonthong, *Data Mining*, 1st Ed., Chamchuree Products, Bangkok, 2015.

[5] A. Chutipascharoen and C. Sanrach, "A comparison of the efficiency of algorithms and feature selection methods for predicting the success of personal overseas money transfer," *KKU Research Journal of Humanities and Social Sciences*, vol. 6, no. 3, pp. 105–113, 2018.

[6] E. Pacharawongsakda, *An Introduction to Data Mining Techniques*, Asia Digital Press, Bangkok, 2014.

[7] O. Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Springer, Berlin, Germany, 2013.

[8] D. A. Pisner and D. M. Schnyer, *Machine Learning*, Academic Press, Massachusetts, United States, 2020.

[9] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63–90, 1993.

[10] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. P. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*, SPSS, Copenhagen, Denmark, 2000.

[11] C. Romero, S. Ventura, and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education*, vol. 51, no. 1, pp. 368–384, 2010.

[12] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Predicting students' perf ormance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.

[13] V. Nakhipova, Y. Kerimbekov, Z. Umarova, L. S. Botayeva, I. Almira, and N. Zhumatayev, "Use of the Naive Bayes classifier algorithm in machine learning for student performance prediction," *International Journal of Information and Education Technology*, vol. 14, no. 1, pp. 92–98, 2024.

[14] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: A case study," *International Journal of Information and Education Technology*, vol. 6, no. 7, pp. 528–533, 2016.
[15] P. Shayan and M. V. Zaanen, "Predicting student performance from their behavior in learning management systems," *International Journal of Information and Education Technology*, vol. 9, no. 5, pp. 337–341, 2019.
[16] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 950–965, 2009.
[17] A. Tamhane, B. Saha, G. Sakarkar, J. Alstott, J. Srivastava, and P. Agarwal, "Predicting student risks through longitudinal analysis," in *Proc. the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1544–1552.
[18] A. Merceron and K. Yacef, "TADA-Ed for educational data mining," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, vol. 7, no. 1, pp. 1–20, 2005.
[19] M. Cocea and S. Weibelzahl, "Log file analysis for disengagement detection in e-Learning environments," *User Modeling and User-Adapted Interaction*, vol. 19, no. 4, pp. 341–385, 2009.
[20] S. B. Shum and R. Ferguson, "Social learning analytics," *Educational Technology & Society*, vol. 15, no. 3, pp. 3–26, 2012.
[21] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an "early warning system" for educators: A proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 588–599, 2010.
[22] S. Amershi and C. Conati, "Automatic recognition of learner groups in exploratory learning environments," in *Proc. the 14th International Conference on Artificial Intelligence in Education*, 2009, pp. 54–61.
[23] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Educational Technology & Society*, vol. 17, no. 4, pp. 49-64, 2014.
[24] J. Peterson, A. Smith, and L. Johnson, "Predicting future professional fields of high school students using machine learning algorithms," *Journal of Educational Data Science*, vol. 10, no. 2, pp. 123–145, 2019.
[25] M. Zhang, W. Lee, and H. Kim, "Career trajectory prediction for university graduates using neural networks and KNN," *International Journal of Career Development*, vol. 15, no. 3, pp. 256–270, 2020.
[26] R. Singh, P. Gupta, and S. Kumar, "Machine learning models for predicting professional fields of engineering students," *Engineering Education Journal*, vol. 18, no. 4, pp. 305–321, 2021.
[27] P. Jaruteerapan, "Factors influencing on choosing studying programs in bachelor's degree level in the Faculty of Management Science, Loei Rajabhat University," *Research and Development Journal, Loei Rajabhat University*, vol. 10, no. 32, pp. 35–46, 2015.
[28] P. Wongthong, W. Kankaew, A. Kwankaew, and Y. Chomdeang, *Applied of Data Mining Techniques for Searching Characteristics of Computer Career*, Research Project, Rajamangala University of Technology Suvarnabhumi, 2019.
[29] W. Panyawong. "Utilization of rule-based predicting models faculty of architecture, urban design and creative arts students at Mahasarakham University by Data Mining," *Journal of Architecture, Design, and Construction*, vol. 2, no. 2, pp. 109–119, 2020.
[30] T. Wannaprapha. "Using data mining techniques for screening people successfully in undergraduate studies of educational technology," M.S. thesis, Dhurakij Pundit University, 2021.
[31] J. Jareanying, "The prediction of student performance using data mining techniques with rapid miner," M.S. thesis, Srinakharinwirot University, 2020.
[32] D. Miao, Q. Duan, H. Zhang, and N. Jiao, "Rough set based hybrid algorithm for text classification," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9168–9174, 2009.