Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPLEMENTATION OF DATA MINING TO DETERMINE STUDENT MAJORS USING THE MACHINE LEARNING

[1]BERNADETE DETA, [2] TUGA MAURITSIUS

[1]Master of Information Systems Management Bina Nusantara University, Indonesia
E-mail: [1]bernadete.deta@binus.ac.id, [2]tuga.mauritsius@binus.edu

ABSTRACT

Implementation of evaluation, planning, and decision making can be done better if an organization has complete, fast, precise, and accurate information. The required information can be retrieved from operational data stored in the integrated database. Data mining has an important role in everyday life. By understanding its definitions, functions, methods and applications, it will be easier to put them into practice. In addition, data collection is also very much needed in various fields of life ranging from telecommunications, insurance, sports, finance, academic fields and other fields. In-depth understanding of data mining is needed to simplify the work. This study examines the extraction of operational data and then analyzes the data using data mining techniques. Data Mining is the process of analyzing data using software. To find certain patterns or rules from a large amount of data that is expected to find knowledge to support decisions. This study uses student data which includes data on National Examination scores, written test scores, interest and aptitude test scores to determine student majors. In this study, the data mining technique used is Classification. The way to do the classification is by using data mining techniques using machine learning. In this study, the data mining technique used is classification using the CRISP-DM method and modeling by comparing the four models namely Decision Tree, Naïve Bayes, KNN Classification and Random Forest with Rapidminer tools to help find characteristics or variables that support in determining student abilities. . Furthermore, it can be used for future student majors. From the results of the analysis that has been done, the model using Decision tree has an accuracy of 95.58%, Naïve Bayes 88.97%, KNN Classification 93.54% and Random forest has an accuracy of 96.46%. The final conclusion is that modeling with Random Forest can be used to help determine student majors in the best private catholic high school on the island of Flores.

Keywords: *Classification, Data Mining, CRISP-DM, Student Majors, Decision Tree Algorithm, Naïve Bayes, KNN Classification, Random Forest, Mechine.*

1. INTRODUCTION

Accurate information is needed in everyday life, information will be an important element in the development of society today and in the future. Utilization of existing data in information systems to support decision-making activities. It is not enough just to rely on operational data, but data analysis is needed to explore the potential of existing information [1] Decision makers try to take advantage of the data warehouse they already have in making decisions. a branch of science to solve the problem of extracting important and interesting information or patterns from large amounts of data, which is called data mining. [2] The use of data mining techniques is expected to provide previous knowledge hidden in the data warehouse so that it becomes valuable information. At any current level, any educational institution, including high school, is required to have a competitive advantage by utilizing all available resources. Information systems are used to obtain, process and disseminate information as well as to support daily operational activities as well as to support strategic decision-making activities. [3] The utilization of departmental data in the data warehouse is currently not optimally and efficiently utilized. student majors have not been fully seen easily and quickly. To see and be able to know the predictions of students' majors, they can maximize and utilize the data that has accumulated in the data warehouse, especially the majors data. [4] By utilizing data mining

techniques, researchers try to explore and obtain useful information to support management decision making, especially those relating to predictions of student majors. [5]The data needed in this study are
1. National exam score (Nes) data (Nes Natural science, Natural science Mathematics, Natural science English, UN Indonesian and average UN)
2. Written test score data (Science written test, Mathematics written test, English written test, Indonesian written test and average written test)
3. Interest and Talent Test Data (Science Interest Interest Test, Mathematics Talent Interest Test, English Talent Interest Test, Indonesian Language Interest Test and Average Talent Interest Test)
This stage includes data selection and grouping the selected attributes or fields into one table, which can be seen in the following two tables.

Table 1.1 Student Mayors Assessment Data

NO	NAMA	Gender	Recommendation	Nasional Exam sroes						Written test						Interst & Talent Test					Leter	Major
			C. Junior high school	Natural Science	Mathematics	Enggilis	Indonesian	Avarage	Lv	NS	Math	Engg	Indo	Avarage	LV	NS	Math	Engg	Indo	Avarage	Value	
1	Prospective student 1	P	Yes	86	87	85	85	85.8	Good	80	86	80	85	84.4	Enough	88	86	86	85	86.3	Good	IPA
2	Prospective student 2	P	Yes	86	87	87	86	86.5	Good	86	80	80	86	84.9	Enough	86	86	85	85	85.5	Good	IPA
3	Prospective student 3	P	No	87	85	85	87	86.0	Good	86	80	80	86	84.7	Enough	86	87	87	85	86.3	Good	IPA
4	Prospective student 4	L	Yes	85	86	85	85	85.3	Good	86	80	80	86	84.3	Enough	88	87	85	87	86.8	Good	IPA
5	Prospective student 5	L	No	88	87	86	85	86.5	Good	86	80	80	86	84.9	Enough	85	84	85	87	85.3	Enough	IPA
6	Prospective student 6	L	No	86	87	86	86	86.3	Good	86	80	80	85	84.7	Enough	85	84	85	87	85.3	Enough	IPA
7	Prospective student 7	L	No	86	87	86	86	86.3	Good	86	80	80	85	84.7	Enough	84	84	85	87	85.0	Enough	IPA
8	Prospective student 8	L	Yes	85	87	86	86	86.0	Good	85	85	86	87	85.9	Good	85	80	85	87	84.3	Enough	IPA
9	Prospective student 9	L	No	86	86	86	86	86.0	Good	85	85	86	87	85.9	Good	85	84	85	87	85.3	Enough	IPA
10	Prospective student 10	P	No	87	86	85	86	86.0	Good	85	85	86	87	85.9	Good	85	84	85	87	85.3	Enough	IPA
11	Prospective student 11	P	Yes	88	87	85	85	86.3	Good	85	85	86	87	86.0	Good	85	84	85	87	85.3	Enough	IPA
12	Prospective student 12	P	No	86	87	87	85	86.3	Good	85	85	85	87	85.9	Good	84	84	85	87	85.0	Enough	IPA
13	Prospective student 13	L	Yes	85	87	87	85	86.0	Good	86	85	85	85	85.7	Good	85	84	85	87	85.3	Enough	IPA
14	Prospective student 14	P	Yes	87	87	87	87	87.0	Good	86	86	85	85	86.3	Good	85	84	85	87	85.3	Enough	IPA
15	Prospective student 15	L	Yes	87	87	87	87	87.0	Good	86	86	87	86	86.7	Good	85	84	85	87	85.3	Enough	IPA
16	Prospective student 16	P	Yes	86	86	87	87	86.5	Good	87	86	87	86	86.5	Good	85	84	85	87	85.3	Enough	IPA
17	Prospective student 17	P	Yes	85	86	85	87	85.8	Good	87	85	87	86	86.0	Good	85	84	85	87	85.3	Enough	IPA
18	Prospective student 18	P	Yes	85	86	85	87	85.8	Good	86	80	80	86	84.5	Enough	85	84	85	87	85.3	Enough	IPA
19	Prospective student 19	L	Yes	85	85	86	85	85.3	Good	86	80	80	86	84.3	Enough	85	84	85	87	85.3	Enough	IPA
20	Prospective student 20	L	Yes	87	85	86	85	85.8	Good	86	80	80	86	84.5	Enough	85	84	85	87	85.3	Enough	IPA

Source: Processed school data 2020.

Information:
- Good = 85 – 100
- Enough = 65 – 84.9
- Less = 0 – 64.9

Based on the background described above, the problems that can be formulated in this study are: "How to apply data mining techniques to determine student majors with Machine Learning using the Decision Tree, Naïve Bayes, KNN Classification and Random Forest algorithm methods in high school on the island of Flores, Indonesia. With the aim of finding/modeling the right algorithm in classifying students' majors, so that students can choose the right major according to their academic abilities, interests, talents, and graduate on time with optimal results.

2. LITERATURE REVIEW

Some of the studies that are used as references in the research that the author conducts include research conducted by Raditya entitled "Implementation of Data Mining Classification to find rain prediction patterns using the C4.5 Algorithm". This research uses the Java programming language and MySQL DBMS to build the application. The accuracy of the prediction pattern obtained can reach 79%. Accuracy is obtained from trials using 2007 weather data as training data and 2008 and 2009 weather data as test data. The next research that becomes a reference is the research conducted by Azimah and Sucahyo with the title "Use of Data Warehouse and Data Mining for Academic Data", [6] The purpose of this research is to find out that evaluation, planning, and decision-making activities will be better if an organization has complete, fast, precise, and accurate information. The next research that becomes a reference is Multiple Intelligences and Reading Comprehension of High School Students: Evaluation of Responses through Educational Data Mining Techniques [7] This study predicts the accuracy of student responses in the actual evaluation, which was conducted in San Jose, Dinagat Islands Regency, Philippines, in determining Multiple Intelligences (MI) and Reading Comprehension in Literature for high school students as the basis for intervention programs. The use of the Naïve Baye algorithm describes an accuracy of 79.93% when applied to the evaluation dataset when performed using the 10-fold cross-validation scheme in the WEKA

software. [8] The following research is used as a reference with the title Early detection of students who have the potential to experience difficulties [9] This study uses data mining methods, this paper presents a new way to identify the profile of new students who may face major difficulties in completing their first studies. academic year. [10] The purpose of this study was to detect potential failures early on using student data available at enrollment, namely school records and environmental factors, with the aim of timely and efficient improvement and/or reorientation of studies. We adapted three data mining methods, namely random forest, logistic regression, and artificial neural network. We designed the algorithm to improve prediction accuracy when multiple classes are in demand. This algorithm is independent of context and can be used in various fields. Real data relating to students at the University of Liège (Belgium), illustrates our methodology. [11]

Follow-up research by Carlyn entitled Analysis of Twitter Sentiment Against the COVID-19 Vaccine in the Philippines Using Naïve Bayes. [12] [13]Aims to collect data on Filipino sentiment regarding the Philippine government's efforts to use the social networking site Twitter. Natural language processing techniques are applied to understand common sentiments, which can assist governments in analyzing their responses. Sentiments were annotated and trained using the Naïve Bayes model to classify English and Filipino tweets into positive, neutral, and negative polarities via data science software RapidMiner. The results yield an accuracy of 81.77%, which exceeds the accuracy of a recent sentiment analysis study using Twitter data from the Philippines. [8] The next research reference is Machine learning compensates for flip-change methods and highlights oxidative phosphorylation in the brain transcriptome of Alzheimer's disease. drug research and development in AD. Briefly, excision of APP by - and -secretase yields 40 and 42 amino Aβ monomers, respectively, which in turn accumulate into amyloid fibrils and cause downstream tau hyperphosphorylation and neurotoxicity, under conditions of insufficient Aβ degradation. The application of ML in AD is focused on the diagnosis of AD from neuroimaging 4. Despite the fact that several AD biologic data have emerged, including genome profiling and electronic health records, a comprehensive understanding of the mechanisms of AD ML has so far not been carried out. realized, mainly due to the lack of the required data density5. We have previously identified MMP14 and dystonin that have the potential to modulate crosstalk. between diabetes and AD by a meta-analysis 6,7. In this study, we applied ML to a publicly available transcriptome dataset from postmortem AD to uncover complex genetic networks and compared the results with conventional fold-change (FC) methods. [14]

Based on some of the previous studies above, it can be said that data mining is a process of extracting data to find important patterns that can be useful information, especially for business owners. For example, finding patterns of consumer behavior from a collection of consumer data over a certain period of time. Just like Data Mining, Machine Learning is also a part of Data Science. Machine Learning is used so that the computer system can carry out the learning process automatically without being given programming instructions first and can increase the accuracy of the prediction results. This is what underlies the author to analyze these data and later can be used as a tool for the easier, faster and more accurate classification process of student majors. From the description above, the author is interested in conducting research with the title "Implementation of Data Mining to Determine Student Majors Using Machine Learning at the best private catholic high school on the island of Flores.".

2.1 Majoring Process in High School

The process of majoring in high school students is carried out directly at the beginning of entering high school or in class X. Currently using the 2013 National Curriculum as a learning guide and Law No. 20 of 2013 as a reference for placement of program specializations. (Kemendikbud, 2013). The criteria for the program majors are carried out based on academic scores, namely the National Examination (UN) SMP scores, written test scores and Interest and Talent Test scores organized by the school. Students who enter class X (ten) will take certain programs, namely: Natural Sciences Program (IPA), Social Science Program (IPS) and Language Program.

2.2 Machine Learning

Machine learning can be defined as the application of computers and mathematical algorithms adopted by means of learning that derives from data and generates predictions of the future. The learning process in question is an effort to acquire intelligence through two stages, including training and testing. [1] [13]

The field of machine learning deals with the question of how to build computer programs to improve automatically based on experience. Recent research reveals that machine learning is divided into three categories: Supervised Learning, Unsupervised Learning, Reinforced Learning. [16]

The technique used by Supervised Learning is a classification method in which the data set is labeled complete to classify the unknown class. While the Unsupervised Learning technique is often called a cluster because there is no need for labeling the data set and the results do not identify examples in the specified class. While Reinforcement Learning is usually between Supervised Learning and Unsupervised Learning This technique works in a dynamic environment where the concept must complete a goal without explicit notification from the computer if the goal has been achieved.[10] [11]

The supervised learning method is based on a collection of labeled data samples. The sample set is used to summarize the characteristics of the behavioral measure distribution across each type of application so as to form a behavioral model from the data. Supervised learning is further grouped into classification and regression problems. A classification problem is when the output variable is categorical, such as red or blue or disease and no disease. [17]While the regression problem is when the output variable is a real value, such as dollars or weights. Supervised learning has several popular algorithms such as Back-propagation, Linear regression, Random Forest, Support Vector Machines, Naive Bayesian, Rocchio Method, Decision Tree, k-Nearest Neighbor, Neural Network, Logistic Regression, and Neural Network. Then several algorithms for classification are mentioned such as support vector machines, Normal Bayesian Classifier, K-Nearest Neighbor, Gradient Boosted Trees, Random Trees, and Artificial Neural Networks. [12] [18] Some of the issues in this category revolve around classification, for example in traffic areas such as the development of Automatic Plate Recognition which can be used in many applications, such as road traffic monitoring, automated toll payments and parking management. Even the use of machine learning in industry is carried out in research. In addition, in the field of medicine such as medical imaging problems, patient data management and gait examination of a person can be predicted with a fairly high accuracy. In the field of technology, such as multimedia text classification, smart watches, etc., of course, many have used machine learning in the supervised learning category. [13]

In the Unsupervised Learning type of learning, the system is equipped with some sample input but no output. Since no output is desired here, categorization is performed so that the algorithm correctly distinguishes between data sets. It is the task of defining a function to describe the hidden structure of unlabeled data. Unsupervised learning is further grouped into clustering and association problems. The clustering problem is where to find the clusters attached to the data, such as grouping customers by purchasing behavior. While the association problem is a rule that describes most of the existing data, such as people who buy A also tend to buy B.[8] Unsupervised learning has several popular algorithms such as k-means, Apriori Independent Subspace Analysis.

Some problems, for example in the financial sector, to review large amounts of data, unsupervised learning can usually be used in the industrial sector, for example in the medical field, unsupervised learning is used in the process of segmenting blood vessels, and technology. such as computer networks and security attack prevention also use this category. Reinforcement learning comes from animal learning theory. This learning requires no prior knowledge, can independently acquire optional policies with knowledge gained through trial and error and continuously interact with a dynamic environment. Reinforcement learning problems are solved by learning new experiences through trial and error Reinforcement learning algorithms are related to dynamic programming algorithms that are often used to solve optimization problems

2.2 Random Forest

Random Forest is a supervised learning algorithm released by Breiman in 2001.[19] [20] Random Forests are commonly used to solve problems related to classification, regression, and so on. There are two things that make this algorithm called random, namely:

1) Each tree grows on a different bootstrap sample taken from the training data randomly.
2) In each split node during decision tree formation, a sample portion of the variables is selected from the original data set and then the best one will be used in that node. This algorithm is a combination of several tree predictors or can be called decision trees where each tree depends on the random vector value which is sampled freely and evenly on all trees in the forest. Prediction results from Random Forest are obtained through the highest results from each individual decision tree (voting for classification and average for regression). For RF consisting of N trees it is formulated as:

$$l(y) = argmax_c(\sum_{n=1}^{N} I_{h_n(y)=c})$$

Where I is an indicator function and is the nth tree of RF. [4] Random Forest has an internal mechanism that provides an estimate of its own generalization error called the out-of-bag (OOB) error estimate. In tree formation only 2/3 of the original data is used in bootstrap sampling. While the remaining 1/3 are classified by the tree that is formed and used to test its performance. OOB error estimation is the average of the prediction errors for each training case y using a tree that does not include y in the bootstrap sample. Then, when the RF is generated, all training cases go through each tree and the proximity matrix of each case is calculated based on the pair of cases that arrive at the same terminal node. [21]

Many studies have proven that Random Forest has good predictive performance in regression and classification in various fields such as financial prediction, remote sensing, as well as genetic and biomedical analysis. RF also shows better performance when compared to other methods such as partial least squares regression, support vector machines and neural networks. [4]

Random Forest has several advantages, namely it can increase the accuracy of the results if there is missing data, and for resisting outliers, as well as efficient for storing data. In addition, Random Forest has a feature selection process which is able to take the best features so that it can improve the performance of the classification model. With feature selection, of course Random Forest can work on big data with complex parameters effectively. Devella.2020.

2.3 Data Mining

Data Mining focuses on developing knowledge form data sets using machine learning techniques. It is also the application of special algorithms to extract patterns from data and convert them into usable information in different domains. (CRIPS DM 1). According to Turban in his book entitled "Decision Support Systems and Intelligent Systems", data mining is a term used to describe the discovery of knowledge in databases. [15] [10] Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from various large databases.

Data mining is the process of discovering information from large and previously unknown data sets. The characteristics of data mining as follows:

1) Data mining is concerned with the discovery of something hidden and certain previously unknown data patterns.
2) Data mining usually uses very large data. Usually big data is used to make the results more reliable.
3) Data mining is useful for making important decisions, especially in strategy.

In its application, data mining is actually a part of Knowledge Discovery in Database (KDD). a process whose job is to extract patterns or models from data using certain algorithms. The KDD process is as follows:

1) Data Selection: data selection from a set of operational data needs to be done before the information mining stage in KDD begins.
2) Preprocessing: before the data mining process can be carried out, it is necessary to carry out a cleaning process with the aim of eliminating data duplication, checking inconsistent data, and correcting errors in data, such as typos. An enrichment process is also carried out, namely the process of "enriching" existing data with other relevant data or information needed for KDD, such as external data or information.
3) Transformation: the process of coding on the data that has been selected, so that the data is suitable for the data mining process. The coding process at KDD is a creative process and is highly dependent on the type or pattern of information to be searched in the database.
4) Data mining: the process of finding interesting patterns or information in selected data using certain techniques or methods.
5) Interpretation/Evaluation: the pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties. This stage is part of the KDD process called interpretation. This stage includes checking whether the patterns or information found contradict the facts or pre-existing hypotheses or not

2.4 RapidMiner

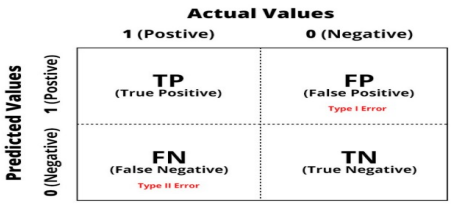
Rapid Miner is software that is open (open source). Rapid Miner is a solution for analyzing data mining, text mining and predictive analytics. Rapid Miner uses a variety of descriptive and predictive techniques to provide users with insights so they can make the best decisions.[16] Rapid Miner has around 500 data mining operators, including

operators for input, output, data preprocessing and visualization. Rapid Miner is a standalone software for data analysis and as
data mining engine that can be integrated into its own product. [2] Rapid Miner is written in Java so it can work on all operating systems. Some of Rapid Miner's features include:

1) Number of data mining algorithms, such as decision trees and self-organization maps.
2) Sophisticated graphic forms, such as overlapping histogram diagrams, tree charts and 3D Scatter plots.
3) Various kinds of plugins, such as text plugins to perform text analysis.
4) Provide data mining and machine learning procedures including: ETL (extraction, transformation, loading), data preprocessing, visualization, modeling and evaluation.
5) The data mining process consists of nestable operators, described in XML, and created with a GUI.
6) Integrating Rapid miner and R.statistics data mining projects

2.5 Confusion matrix

The confusion matrix is also known as the error matrix. Basically the confusion matrix provides information on the comparison of the classification results performed by the system (model) with the actual classification results. The confusion matrix is in the form of a matrix table that describes the performance of the classification model on a series of test data whose actual values are known. The picture below is a confusion matrix with 4 different combinations of predicted values and actual values. It can look like the image below. There are 4 terms that represent the results of the classification process in the confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).



(source: https://ksnugroho.medium.com/confusion-matrix)

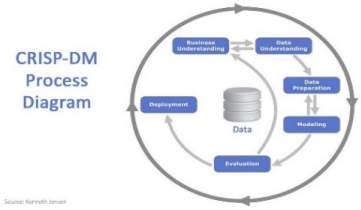
The benefits of the confusion matrix are:

1) Shows how the model makes predictions.
2) It not only provides information about the errors made by the model but also the types of errors made.
3) Each column of the confusion matrix represents an instance of the prediction class.
4) Each row of the confusion matrix represents an instance of the actual class.

How to measure performance metrics from the confusion matrix can use the confusion matrix to calculate various performance metrics to measure the performance of the model that has been created. Some of the popular performance metrics that are commonly and frequently used are accuracy, precision, and recall.

3. RESEARCH METHODOLOGY

The research method used in this study follows the stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM) model. The stages of CRISP-DM,[17] are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (CRISP, 2009). This research requires the following stages:



Picture 3.1 CRIPS-DM Proses Diagram

Explanation of the CRISP_DM Process Diagram, is as follows:

1) Business Understanding, is an understanding of the substance of the data mining activities to be carried out, the needs from a business perspective. Its activities include determining business goals or objectives, understanding business situations, translating business objectives into data mining objectives.
2) Data Understanding, namely collecting data, studying data to be able to understand the data to be used in research, identifying problems related to data.
3) Data Preparation, at this stage a database structure will be prepared to facilitate the mining process.
4) Modeling stage, is the stage of determining the data mining technique used, determining data mining tools, data mining algorithms, determining parameters with optimal values.
5) Evaluation, is the stage of evaluating whether the data mining modeling has met the research objectives that have been determined at the business understanding stage, namely the

Business Understanding Phase.

6) Deployment is the manufacturing stage report of research results, reporting of results can be done after completing the evaluation of the grouping mode.[18]

The Cross Industry Standard Process for Data Mining (CRISPDM) has become the standard for organizing and conducting data mining projects. It is considered a top methodology for data mining, or data science projects. Some researchers place more emphasis on the role of effective project management and especially systematic documentation as key success factors for Knowledge Discovery in Databases (KDD) projects. Despite all the efforts made to introduce various methods for managing data mining projects, others argue that some of the common pitfalls that occur in DM projects can be summarized as a lack of methodology for project development. Explanation of the CRISP_DM Process Diagram, is as follows:

1) Business Understanding, is an understanding of the substance of the data mining activities to be carried out, the needs from a business perspective. Its activities include determining business goals or objectives, understanding business situations, translating business objectives into data mining objectives.
2) Data Understanding, namely collecting data, studying data to be able to understand the data to be used in research, identifying problems related to data.

3) Data Preparation, at this stage a database structure will be prepared to facilitate the mining process.

4) Modeling stage, is the stage of determining the data mining technique used, determining data mining tools, data mining algorithms, determining parameters with optimal values.

5) Evaluation, is the stage of evaluating whether the data mining modeling has met the research objectives that have been determined at the business understanding stage, namely the Business Understanding Phase.

6) Deployment is the manufacturing stage report of research results, reporting of results can be done after completing the evaluation of the grouping mode.[18]

The Cross Industry Standard Process for Data Mining (CRISPDM) has become the standard for organizing and conducting data mining projects. It is considered a top methodology for data mining, or data science projects. Some researchers place more emphasis on the role of effective project management and especially systematic documentation as key success factors for Knowledge Discovery in Databases (KDD) projects. Despite all the efforts made to introduce various methods for managing data mining projects, others argue that some of the common pitfalls that occur in DM projects can be summarized as a lack of methodology for project development.

4. RESULTS AND DISCUSSION

4. 1 Business Understanding

The implementation of data mining in this study is directly related to data on majors for high school students. As well as to see what parameters affect the majors that have an impact on changing majors by class X students in the middle of the semester.

Next, explore knowledge about how to model these majors with machine learning with the Decision Tree, Naïve Bayes, KNN Classification, and Random forest algorithms, so that the high school, especially the academic part of the curriculum, can understand and use the modeling in determining majors for students.

Using machine learning the level of accuracy is better. Machine Learning can also help academics to make it easy, fast, and precise in determining the majors for class X students at the beginning of school. Determining the right major can help students in following the lessons well according to their interests, talents and academic abilities.

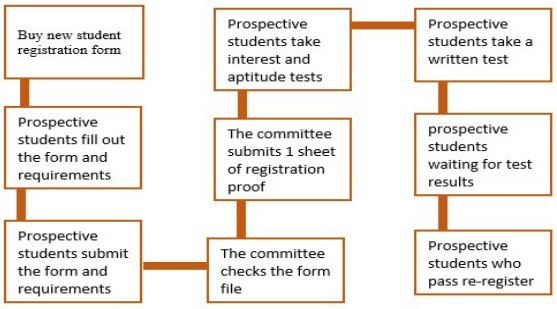
Business understanding in schools in the acceptance of prospective students to specialization or determination of specialization programs are as follows:

a) Prospective students buy registration new student admission form committee
b) Fill in the form and complete the requirements that have been set
c) Submit the completed form and its requirements to the committee
d) The committee checks the completeness of the requirements and records prospective student data in the registration book.
e) Submit 1 sheet of proof of registration to prospective students who have been given a registration number.
f) Prospective students take an interest and aptitude test, the results contain recommendations for majors that are in accordance with the interests and talents of prospective students.
g) Take a written test.

The above process can be seen as shown below:

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



Picture 4.1 Business Understanding in Schools

4.2 Data Understanding

This study uses data from the best private catholic high school senior high school students on the island of Flores for the 2020/2021 academic year. Data were obtained from the Deputy Principal of the High School Curriculum, which consisted of 376 student data. From the data, the data analysis process is carried out so that data in the form of majors data tables are obtained, which are ready for data mining. The data collected from student profiles with attributes for majors are; Name of Student, Gender, Origin of school, Recommendation for Junior High School Counseling Guidance teacher, National Examination Score (Nasional exam Mathematics Nes, English UN Indonesian Language National Examination, National Examination Average), Written test (Ns, Mathematics WT, English WT, Bindo WT , Average written test), Interest and Talent Test (Science, Mathematics Tts, English Tts, and Indonesian Tts) and Department as a class/decision label. The label is the research target variable containing the majors, namely "Natural Science", "Social science", and "Language". The following is a table of student data for majors.

Table 4.1 Data Training

NO	NAMA	Gender	Recommendation (L.Junior high school)	Nasional Exam snores					Lv	Written test					LV	Interest & Talent Test					Leter Value	Majors
				Natural Science	Mathematics	Enggilis	Indonesian	Average		NS	Math	Engg	Indo	Average		NS	Math	Engg	Indo	Average		
1	Prospective student 1	P	Yes	86	87	85	85	85.8	Good	80	86	80	86	84.4	Enoungh	80	86	86	85	86.3	Good	IPA
2	Prospective student 2	P	Yes	86	87	87	86	86.5	Good	86	80	80	86	84.9	Enoungh	86	86	85	85	85.5	Good	IPA
3	Prospective student 3	P	No	87	85	85	87	86.8	Good	86	80	80	86	84.7	Enoungh	85	87	87	85	86.3	Good	IPA
4	Prospective student 4	L	Yes	85	86	86	85	85.3	Good	86	80	80	86	84.3	Enoungh	80	87	85	87	86.8	Good	IPA
5	Prospective student 5	L	No	86	87	86	85	86.5	Good	86	80	80	86	84.9	Enoungh	85	84	85	87	85.3	Enoungh	IPA
6	Prospective student 6	L	No	86	87	86	86	86.3	Good	86	80	80	86	84.7	Enoungh	85	84	85	87	85.3	Enoungh	IPA
7	Prospective student 7	L	No	86	87	86	86	86.3	Good	86	80	80	86	84.7	Enoungh	84	84	85	87	85.3	Enoungh	IPA
8	Prospective student 8	L	Yes	85	87	86	86	86.3	Good	85	85	86	87	85.8	Good	85	80	85	87	84.3	Enoungh	IPA
9	Prospective student 9	L	No	86	86	86	86	86.3	Good	85	85	86	87	85.8	Good	85	84	85	87	85.3	Enoungh	IPA
10	Prospective student 10	P	No	87	86	85	86	86.8	Good	85	85	86	87	85.8	Good	85	84	85	87	85.3	Enoungh	IPA
11	Prospective student 11	P	Yes	88	87	85	85	86.3	Good	85	85	86	87	86.8	Good	85	84	85	87	85.3	Enoungh	IPA
12	Prospective student 12	P	No	86	87	87	85	86.3	Good	85	85	85	87	85.8	Good	84	84	85	87	85.8	Enoungh	IPA
13	Prospective student 13	L	Yes	85	87	87	85	86.8	Good	86	85	85	86	85.7	Good	85	84	85	87	85.3	Enoungh	IPA
14	Prospective student 14	P	Yes	87	87	87	87	87.3	Good	86	86	85	85	86.3	Good	85	84	85	87	85.3	Enoungh	IPA
15	Prospective student 15	L	Yes	87	87	87	87	87.3	Good	86	86	87	86	86.3	Good	85	84	85	87	85.3	Enoungh	IPA
16	Prospective student 16	P	Yes	86	86	87	87	86.5	Good	87	86	87	86	86.5	Good	85	84	85	87	85.3	Enoungh	IPA
17	Prospective student 17	P	Yes	85	86	85	87	85.8	Good	87	85	87	86	86.8	Good	85	84	85	87	85.3	Enoungh	IPA
18	Prospective student 18	P	Yes	85	86	85	87	85.8	Good	86	86	86	86	84.5	Enoungh	85	84	85	87	85.3	Enoungh	IPA
19	Prospective student 19	L	Yes	85	85	86	85	85.3	Good	86	80	80	86	84.3	Enoungh	85	84	85	87	85.3	Enoungh	IPA
20	Prospective student 20	L	Yes	87	86	86	85	86.8	Good	86	80	80	86	84.3	Enoungh	85	84	85	87	85.3	Enoungh	IPA

(School archive source 2020)

4.3 Data Preparation

The process or steps carried out at this stage are as follows:

a. Data cleaning

Initial data acquisition from school institutions in the form of 3 tables, then the data is combined to produce one file with 376 records and 23 attributes. Next, data cleaning is performed. In this process inconsistent and noise data, namely data on the value of empty letters of interest and talent, are cleaned and/or discarded. So that the data acquisition for testing data is 376 records, 21 attributes including 2 special attributes, namely ID and Label.

b. Data Integration

The findings in this study, the initial data is still separate in the form of Exel files for majors, namely IPA with a total of 86 records, BHS majors with a total of 63 records, Social Studies majors with 227 records and each department has 24 attributes. Then the data is combined so as to produce 376 records with 24 attributes and the data is named Merged data file. This data looks like the following.

Table 4.2 Data Integration

No	NAMA	Gender	Asal SMP	R.BK SMP	UN IPA	UN Mat	UN Bing	UN Bindo	RUN	Ns Run	TT IPA	TT MAt	TT Eng	TT Bindo	RTT	Ns Rtt	Nb IPA	Nb Mat	Nb Bing	Nb Bindo	RNN	Jurusan
1	Agustina A. Wea	P	Boawae	Yes	86	87	85	85	85.8	Bagus	80	86	80	85	84.4	Cukup	86	86	85	85	86.3	IPA
2	Zakarias Bala Mitan	L	Boawae	Yes	80	86	86	87	84.8	Cukup	87	86	87	85	85.6	Baik	86	87	85	86	86.0	IHS
3	Zakarias Bala M.	L	Boawae	No	87	87	85	86	86.3	Bagus	85	85	87	85	85.9	Bagus	87	87	87	86	86.8	IPA
4	Yustina Moi Logo	P	Luar Boawae	Yes	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
5	Yustina Gowa	P	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
6	Yunita Tepo	P	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
7	Yunior Habibly Haji Sulaiman	L	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
8	Yulius S. Wasio Keo	L	Boawae	Yes	76	80	86	87	82.3	Cukup	88	86	86	89	84.7	Baik	86	87	85	86	86.0	IHS
9	Yulianus Jato	L	Boawae	No	87	87	88	86	87.0	Bagus	83	80	85	85	85.3	Cukup	87	87	86	85	86.3	IPA
10	Yulianus Jato	L	Boawae	No	76	75	86	87	81.0	Cukup	88	88	88	89	84.2	Baik	86	87	87	86	86.5	IHS
11	Yuliana Coo Sanda	P	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
12	Yuliana Bengu	P	Luar Boawae	No	76	75	87	85	80.8	Cukup	88	88	88	89	84.1	Baik	86	87	87	86	86.5	IHS
13	Yovita Bupu	P	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
14	Yosef Adriano Pusiima Dopo	L	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
15	Yonianus Beko	P	Luar Boawae	Yes	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
16	Yolanta Wulu	P	Luar Boawae	Yes	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
17	Yohanes Tie	L	Luar Boawae	Yes	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
18	Yohanes Sebe	L	Luar Boawae	No	76	75	87	85	80.8	Cukup	88	88	88	89	84.1	Baik	86	86	87	86	86.3	IHS
19	Yohanes S. Mite	L	Boawae	No	86	87	85	86	86.0	Bagus	83	80	85	85	84.8	Cukup	86	87	86	85	86.0	IPA
20	Yohanes Oslanus Nete	L	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
21	Yohanes Ole Tunga	L	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
22	Yohanes Meo	L	Luar Boawae	Yes	79	80	87	85	82.8	Cukup	88	88	88	89	85.2	Baik	86	85	85	86	85.5	IHS
23	Yobanes Maria Yassery Wera	L	Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
24	Yohanes Dominsio Raja	L	Luar Boawae	Yes	79	80	87	85	82.8	Cukup	88	88	88	89	85.2	Baik	86	86	85	86	85.8	IHS
25	Yohanes Busa Se	L	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS
26	Yohana Tipa	P	Luar Boawae	No	80	86	86	87	84.8	Cukup	87	86	85	85	85.4	Cukup	86	87	85	86	86.0	IPS

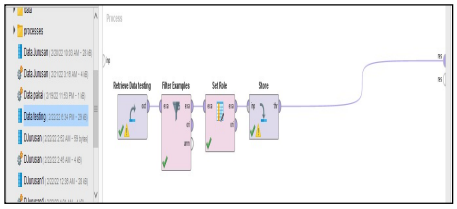
c. Data transformation

At this stage, from the acquisition of data that has been integrated, then this data is converted into a form suitable for data testing.

The data is then imported into Rapidminer. then used for majors modeling with Decision tree, Nave Bayes random fores and KNN Classification. Rapidminier tools used in this research is RapidMiner 9.10. The testing data is then formatted for these data types, bimonial and polymonial and set role labels on the destination data, namely majors. The amount of data is 376 records and 19 attributes and 2 special attributes. (ID and Label). The attributes that are labeled are the attributes of the Department which include Natural science, Language and Social science. The data preparation process on rapidminer looks like the following picture.

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



Picture 4.1 Data Preparation Proses

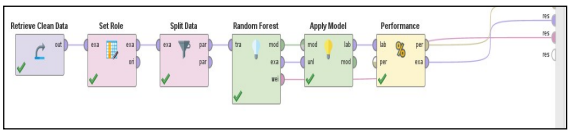
The picture above is a complete data trending and data testing. then the data, filtered and cleaned, is labeled and ID on the role set. Furthermore, the clean data is ready for modeling. This data is stored with the deposit operator.
The results of the above process when executed will look like the following table.

Table 4.3 Data Clean

Gender	Mayor	Rc JNS	NES-NS	UN Mat	NES-Egg	NES Indo	A-NES	LV NES	WT NS	WT-Mat	WT-Egg
P	Indonesian	No	83	86	76	86	82.750	Enough	88	86	90
P	Social science	No	80	80	78	79	79.250	Not enough	75	85	85
L	Social science	No	80	80	78	79	79.250	Not enough	75	85	85
L	Social science	No	80	80	78	79	79.250	Not enough	75	85	85
P	Social science	Yes	80	80	78	79	79.250	Not enough	75	85	86
P	Social science	No	80	80	78	79	79.250	Not enough	80	85	86
L	Indonesian	No	83	86	79	86	83.500	Enough	88	86	90
L	Social science	No	80	80	79	79	79.500	Not enough	82	86	87
L	Social science	No	80	80	79	79	79.500	Not enough	80	85	85
L	Social science	No	80	80	79	79	79.500	Not enough	80	86	85
P	Indonesian	No	75	76	80	87	79.500	Not enough	88	88	87
P	Indonesian	No	75	76	80	87	79.500	Not enough	88	88	87
L	Indonesian	No	75	76	80	87	79.500	Not enough	88	88	87
L	Indonesian	No	75	76	80	87	79.500	Not enough	88	88	87

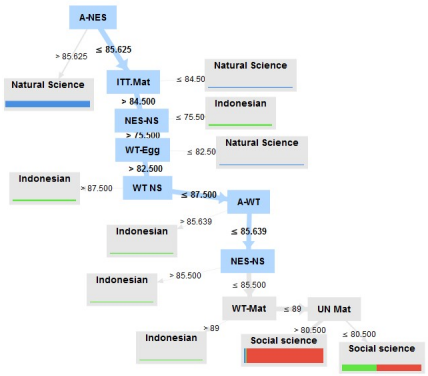
4.4 Modeling

At the modeling stage, namely making predictive models, namely grouping students' majors based on the values obtained by school students. At this stage, you can use statistics and machine learning to gain useful insights from the data to achieve research objectives. To do student data clustering in schools using Machine Learning. Summary of results using Machine Learning with 4 models of Decision treen Naïve Bayes, Random Forest, KNN Classification. by the author displays the random Forest model because random Forest has the highest accuracy. Modeling with random forest is taking clean data that has been stored in the local repository. then the data is labeled using the set role operator. The next step uses the split data operator to divide the data by the ratio: 0.7 and 0.3. There are 2 ports available here using Random Forest modeling which is 70% for training and 30% for applying model. then connected to the next apply mode operator using performance, directly connected and run / run. The stages of the modeling process can be seen as follows:.



Picture 4.2 Process Modeling Random Forest

The results of the random forest pattern above can look like the following image,



Picture 4.3 Tree Random Forest

From the Decision Tree above, it can be seen that the average value of the National Examination has the highest score. Further it can be read as follows:

- If the acquisition value is 85, then it is clear that it is included in NS
- If the average NE score is 85 and the English written test is > 82, the average written test score is > 82 then you enter science,
- If the written average score is >82, National Exam Ns is sufficient, enter the language department.
- If the value of interest in Indonesian talent is 83, the average written test is 85 and the English NE is > 85 then enter the language department
- If the average written test score is > 83 and English > 85 Interests and Talents 86 and NE Natural Science > 79 enter the Social Sciences major. And it can be drawn as follows: From the Decision Tree above, it can be seen that the average value of the National Examination has the highest score. Further it can be read as follows:
- If the acquisition value is 85, then it is clear that it is included in NS
- If the average Nes score is 85 and the English written test is > 82, the average written test score is > 82 then you enter science,

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- ❖ If the written average score is >82, Nes NS is sufficient, enter the language department.
- ❖ If the value of interest in Indonesian talent is 83, the average written test is 85 and the English NE is > 85 then enter the language department
- ❖ If the average written test score is > 83 and English > 85 Interests and Talents 86 and NE NS > 79 enter the Social Sciences major

And it can be drawn as follows:

Tree

```
A-NES > 85.625: Natural Science {Natural Science=76, Indonesian=0, Social science=0}
A-NES ≤ 85.625
|   ITT.Mat > 84.500
|   |   NES-NS > 75.500
|   |   |   WT-Egg > 82.500
|   |   |   |   WT NS > 87.500: Indonesian {Natural Science=0, Indonesian=13, Social science=0}
|   |   |   |   WT NS ≤ 87.500
|   |   |   |   |   A-WT > 85.639: Indonesian {Natural Science=0, Indonesian=3, Social science=0}
|   |   |   |   |   A-WT ≤ 85.639
|   |   |   |   |   |   NES-NS > 85.500: Indonesian {Natural Science=0, Indonesian=1, Social science=0}
|   |   |   |   |   |   NES-NS ≤ 85.500
|   |   |   |   |   |   |   WT-Mat > 89: Indonesian {Natural Science=0, Indonesian=1, Social science=0}
|   |   |   |   |   |   |   WT-Mat ≤ 89
|   |   |   |   |   |   |   |   UN Mat > 80.500: Social science {Natural Science=4, Indonesian=2, Social science=181}
|   |   |   |   |   |   |   |   UN Mat ≤ 80.500: Social science {Natural Science=0, Indonesian=34, Social science=44}
|   |   |   WT-Egg ≤ 82.500: Natural Science {Natural Science=1, Indonesian=0, Social science=0}
|   |   NES-NS ≤ 75.500: Indonesian {Natural Science=0, Indonesian=14, Social science=0}
|   ITT.Mat ≤ 84.500: Natural Science {Natural Science=2, Indonesian=0, Social science=0}
```

Picture 4.4 Description Random Forest

4.5 Evaluation

In the tests carried out with the Random Forest algorithm, the results are as in table 4.4 below.

Table 4.3 Accuracy Random Forest

accuracy: 96.54%

	true Natural Science	true Indonesian	true Social science	class precision
pred. Natural Science	87	0	0	100.00%
pred. Indonesian	0	53	0	100.00%
pred. Social science	3	10	223	94.49%
class recall	96.67%	84.13%	100.00%	

From the table of modeling results with random forest, it can be seen;

a. Accuracy = 96 %

b. Recall: 100% True Natural science recall class prediction, True Language with gain 78.95%, true social science 96,86%

b. Precision: class precision Ns 100%, class precision Language 100%, Class Social science precision 94.44%. Based on the results of modeling with random forest, it produces a very high level of accuracy and has a very good classification level. The description can also look like the following.

d. Weight: The most influencing attribute in determining majors is the National Examination Average (Nea) and the Average Written Test Score (Awts). This can be seen in the following image.

accuracy: 96.54%

	true Natural Science	true Indonesian	true Social science	class precision
pred. Natural Science	87	0	0	100.00%
pred. Indonesian	0	53	0	100.00%
pred. Social science	3	10	223	94.49%
class recall	96.67%	84.13%	100.00%	

accuracy: 96.54%

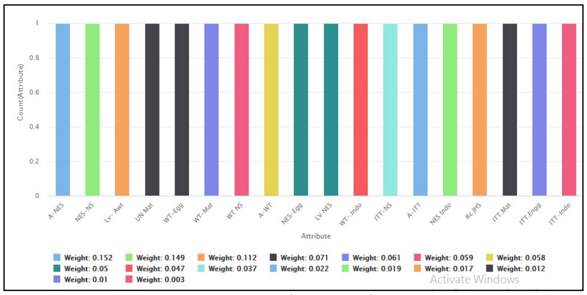
	true Natural Science	true Indonesian	true Social science	class precision
pred. Natural Science	87	0	0	100.00%
pred. Indonesian	0	53	0	100.00%
pred. Social science	3	10	223	94.49%
class recall	96.67%	84.13%	100.00%	

accuracy: 96.54%

	true Natural Science	true Indonesian	true Social science	class precision
pred. Natural Science	87	0	0	100.00%
pred. Indonesian	0	53	0	100.00%
pred. Social science	3	10	223	94.49%
class recall	96.67%	84.13%	100.00%	

attribute	weight
ITT-Indo	0.012
NES Indo	0.030
WT-.Indo	0.043
ITT.Mat	0.014
ITT.Engg	0.014
ITT-NS	0.024
A-ITT	0.034
Rc JHS	0.025
A-WT	0.057
WT NS	0.057
A-NES	0.136
WT-Mat	0.063
NES-NS	0.156
LV NES	0.053
Lv- Awt	0.093

Picture 4.7 Weinght Random Forest Majors



Picture 4.8 Weinght Visualisation Majors

Figure 4.14 above is a graph that shows the weight gain of the most influential attribute in determining majors using the Random Forest model.
Based on the results of modeling with Random Forest, the accuracy is 96.46%. These results are very good accuracy and have a very good classification level.

4.6 Deployment

Based on the test results of the four majors modeling algorithms with the Rapid miner learning tools, the accuracy is very good. Random forest has very good accuracy compared to Decision Tree, Naïve Bayes and KNN Classification. Recommendation to the academic section of the curriculum is to model student majors using the Random forest algorithm. An important step taken by school academics is to integrate all data into the data warehouse. and then using the Random forest algorithm.

5. CONCLUSION AND SUGGESTIONS

a. Conclusion:

The conclusion regarding the majors of prospective students in high school using data mining is that the selection committee of prospective new students can classify prospective new students based on majors using Machine Learning with the attributes used include National Examination Results, Written Tests, and Talent Interest Tests. Then the system of majors and student admissions makes it easier for the committee. Interested prospective students in the best private high schools on the island of flores can use data mining applications using Machine Learning with the Random Forest model. The calculation accuracy that has been done is 94.46%. The use of Machine Learning can facilitate decision making in determining student majors.

b. Suggestion:

From the research that has been done and as the end of this paper, can provide suggestions in the hope that it can be useful for the school in order to facilitate the majors of prospective new students and improve student achievement. Suggestions for further research are analyzing data and other attributes not only from Natonal Exam scores, written test scores and interest and talent test scores and can be added with parental recommendations and psychological tests so that the resulting knowledge can be better.

REFERENCES

[1] M. Yaumi, S. F. S. Sirate, and A. A. Patak, "Investigating Multiple Intelligence-Based Instructions Approach on Performance Improvement of Indonesian Elementary Madrasah Teachers," *SAGE Open*, vol. 8, no. 4, 2018, doi: 10.1177/2158244018809216.

[2] P. Saengsikhiao and J. Taweekun, "The Data Mining Technique Using RapidMiner Software for New Zeotropic Refrigerant," *J. Adv. Res. Fluid Mech. Therm. Sci.*, vol. 83, no. 1, pp. 70–90, 2021, doi: 10.37934/arfmts.83.1.7090.

[3] N. Panjaitan and A. Y. Bin Ali, "Clasification of ergonomics levels for research," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 505, no. 1, 2019, doi: 10.1088/1757-899X/505/1/012040.

[4] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, pp. 36–51, 2018, doi: 10.1016/j.dss.2018.09.001.

[5] A. Raditya, "Omplementasi Data Mining Classification Untuk Mencari Pola Prediksi Hujan Dengan Menggunakan Algoritma C45," 2014.

[6] J. Liu, G. Shi, J. Zhou, and Q. Yao, "Prediction of College Students' Psychological Crisis Based on Data Mining," *Mob. Inf. Syst.*, vol. 2021, 2021, doi: 10.1155/2021/9979770.

[7] L. H. Iwaya, S. Fischer-Hübner, R. M. Åhlfeldt, and L. A. Martucci, "Mobile health systems for community-based primary care: identifying controls and mitigating privacy threats," *JMIR mHealth uHealth*, vol. 7, no. 3, 2019, doi: 10.2196/11642.

[8] N. S. Buot, "International Journal of Advanced Trends in Computer Science and Engineering Multiple Intelligences and Reading Comprehension of Senior High School Students : A Response Evaluation through Educational Data Mining Technique," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 2871–2876, 2019, [Online]. Available: http://files/351/Nenita S. Buot (2019)- MI and reading comprehension of senior high school students.pdf.

[9] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Inf.*, vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[10] A. A. Jamjoom, "The use of knowledge extraction in predicting customer churn in B2B," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00500-3.

[11] A. Santos-Olmo, L. E. Sánchez, D. G. Rosado, E. Fernández-Medina, and M. Piattini, "Applying the action-research method to develop a methodology to reduce the installation and maintenance times of Information Security Management Systems," *Futur. Internet*, vol. 8, no. 3, pp. 1–24, 2016, doi: 10.3390/fi8030036.

[12] N. A. Deraman, A. G. Buja, K. A. F. A. Samah, M. N. H. H. Jono, M. A. M. Isa, and S. Saad, "A social media mining using topic modeling and sentiment analysis on tourism in Malaysia during COVID19," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 704, no. 1, 2021, doi: 10.1088/1755-1315/704/1/012020.

[13] J. Benito-León, M. D. Del Castillo, A. Estirado, R. Ghosh, S. Dubey, and J. I. Serrano, "Using unsupervised machine learning to identify age- And sex-independent severity subgroups among patients with COVID-19: Observational longitudinal study," *J. Med. Internet Res.*, vol. 23, no. 5, pp. 1–14, 2021, doi: 10.2196/25988.

[14] A. M. S. Ridaryanto, Refi Kautsar Firmansyah, Rano Kartono, "International Journal of Advanced Trends in Computer Science and Engineering Available Online at http://www.warse.org/ijatcse/static/pdf/file/ijatcse02422015.pdf," *E3S Web Conf.*, vol. 4, no. 2, pp. 15–21, 2015.

[15] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," *Int. J. Comput. Sci. Manag. Res.*, vol. 1, no. 4, pp. 686–690, 2012.

[16] S. Marzukhi, N. Awang, S. N. Alsagoff, and H. Mohamed, "RapidMiner and Machine Learning Techniques for Classifying Aircraft Data," *J. Phys. Conf. Ser.*, vol. 1997, no. 1, 2021, doi: 10.1088/1742-6596/1997/1/012012.

[17] J. Cheng, H. P. Liu, W. Y. Lin, and F. J. Tsai, "Machine learning compensates fold-change method and highlights oxidative phosphorylation in the brain transcriptome of Alzheimer's disease," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021, doi: 10.1038/s41598-021-93085-z.

[18] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[19] Y. K. Salal, S. M. Abdullaev, and M. Kumar, "Educational data mining: Student performance prediction in academic," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4C, pp. 54–59, 2019.

[20] R. Hasan, S. Palaniappan, S. Mahmood, K. U. Sarker, and A. Abbas, "Modelling and predicting student's academic performance using classification data mining techniques," *Int. J. Bus. Inf. Syst.*, vol. 34, no. 3, pp. 403–422, 2020, doi: 10.1504/IJBIS.2020.108649.

[21] E. Rangaswamy, G. Periyasamy, and N. Nawaz, "A study on singapore's ageing population in the context of eldercare initiatives using machine learning algorithms," *Big Data Cogn. Comput.*, vol. 5, no. 4, 2021, doi: 10.3390/bdcc5040051.