Advances in Computer and Communications, 2020, 1(1), 46-52
http://www.hillpublisher.com/journals/acc/
ISSN Online: ****-****
ISSN Print: ****-****

Evaluation of Student Academic Performance Using Naïve Bayes Classifier

Khin Shin Thant*, Ei Theint Theint Thu, Myat Mon Khaing, Khin Lay Myint, Hlaing Htake Khaung Tin

Faculty of Information Science, University of Computer Studies, Hinthada, Myanmar.

How to cite this paper: Khin Shin Thant, Ei Theint Theint Thu, Myat Mon Khaing, Khin Lay Myint, Hlaing Htake Khaung Tin. (2020) Autopoietic Computing Systems and Triadic Automata: The Theory and Practice. *Advances in Computer and Communication*, 1(1), 46-52.
DOI: 10.26855/acc.2020.12.005

Received: November 20, 2020
Accepted: December 18, 2020
Published: December 24, 2020

***Corresponding author:** Khin Shin Thant, Faculty of Information Science, University of Computer Studies, Hinthada, Myanmar.
Email: hlainghtakekhaungtin@gmail.com

Abstract

The best way to achieve the highest quality in the higher education system is to increase the ability of students to improve their performance. Students' ability to find information is called rules for identifying information. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Data mining is used to find knowledge that may be useful in applying active learning in the technology perspective. There are two approaches which can be used to discover knowledge by data mining techniques such as prediction, classification, clustering, association rule and statistical methods. Among them, this paper predicts using data mining classification method. The data used in this paper is based on the student records that collected from department at the end of each semester. The main purpose of this paper is to assess students' ability to study in quality education and to assess student performance through the Naïve Bayes classification. These findings will be very helpful to teach data mining subject in the coming year for students.

Keywords

Evaluation, Data Mining, Students' Performance, Classification, Naïve Bayes

1. Introduction

Nowadays, data mining techniques have been applied in various applications. These techniques are mostly used in education field to extract the hidden information and discover patterns from educational dataset. Data mining, known as Knowledge Discovery in Databases (KDD), is to discover of useful information from large collections of data and extract patterns of stored data by using various methods and algorithms. There are two approaches: statistical methods and data mining techniques. The data mining techniques are Classification, Clustering, Prediction, and Association rule, Neural Networks, Decision Trees and Nearest Neighbor Method. This paper applies Naïve Bayes Algorithm in Classification to predict the students' performance in academic year. This paper examines the educational domain of students' performance data. The algorithm is to accurately predict the particular subject and make decisions in time [1].

Many researchers have been done researches by using data mining techniques and tools. Using data mining techniques and tools become very popular in different areas, especially educational field. The work processed based on several attributes to predict performance of the students by using classification techniques. Examine the student's weaknesses and strengths that will contribute to future performance [2].

2. Data Mining Techniques

2.1 Classification

Supervised machine learning is the most popular search engine. It is used to classify data based on training data. So it's called the Training Datasets. Pre-supervising or overseeing patterns on target data. Learning is analyzed through a classification algorithm. The test facts were used to guess the correctness of the data cataloguing procedures. If accuracy is good and acceptable, new instructions may apply [3-5].

2.2 Clustering

Clustering is an unregulated machine learning and statistical information investigation technology. It is used to classify the same information as a homogeneous cluster and to run a domain-controlled database of different domains.

2.3 Prediction

Many real problems are not just speculations. As a result, more sophisticated methods (such as logistic regression, decision trees, or neural tubes) could be needed to predict future values. It can be convenient to estimate the regression method. The CART (Classification and Regression Trees) can be used to distinguish deciduous trees for building trees and for predicting variable responses.

2.4 Association Rule

Association rule mining is a very common form of occurrence. Relationships or relations databases; It is intended that you frequently study organizations from a variety of databases such as trading databases and other storage formats. Association Rule Mining is sometimes referred to as "Market Basket Basket Analysis".

2.5 Neural Networks

Normally, the artificial neural network (ANN) is a mathematical or computational model that stimulates the structure and/or function of the neural network. A neural network is made up of a group of interconnected neurons, such as a continuum; It uses the computational method of computing and processes information. Usually, the ANN is an adaptive system that changes its structure based on external information flowing through the network during the lesson. Modern neural networks are non-linear statistical data modeling tools. They are often used to design complex relationships between inputs and outputs or to find patterns in data.

2.6 Decision Trees

This is one of the most popular approaches in knowledge discovery and data mining. Algorithm of Decision Tree is in Data Mining. A decision tree is a supervised learning approach wherein we train the data present with already knowing what the target variable actually is. In Decision Tree, the algorithm splits the dataset into subsets on the basis of the most important or significant attribute. It is flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch denotes an outcome of test, and each leaf node holds a class label. The highest node in a tree is the root node. For the unknown class labels, the tuple, X, and attribute values of the tuple are being tested against the decision tree. A path traverses a leaf node that receives a linear estimate of that tuple. Decision tree is useful because building decision trees does not require any domain knowledge. This can handle views data. The study and classification stages of the decision making process are simple and fast. Representing their knowledge in a tree-based way is easy for users to upload. Decisions tree types determine good accuracy.

2.7 Nearest Neighbor Method

The k-nearest-neighbor (KNN) algorithm is a simple but effective method of estimating the labeling level of a sample probe based on its near information. KNN is one of the easiest and most straightforward data search technologies. This is called Memory-based Classification. Training examples need to be in memory at run-time. When dealing with continuous attributes, use the Euclidean distance between attributes. Calculate K-NN is an example of learning to be lazy or lazy. It is closest to the function locally and defines all calculations until the function is evaluated. A useful way for both classification and retrenchment is to weigh the contributions of neighbors. Only then will the average neighbor be able to support the average over the distances taken by the objects for the class. Item is known for its k-NN classification) or (for - NN decay). This can be supposed of as the preparation set for the

algorithm, though no categorical training step is required.

2.8 Linear Regression

This is prediction technique that predicts a numeric variables including marks, age and weight. A linear regression analysis technique is a data analysis technique. It is used to determine the degree of linear connection between independent variables and more than one independent variable. There are two types of linear regression, simple linear regression and multiple linear regression. A simple linear regression is used to estimate the value of a dependent variable for an independent variable. Two or more independent variables are used to estimate the value of a dependent variable. The difference between the two variables is the number of independent variables. In both cases, there is only one dependent.

2.9 Support Vector Machine

Support vector machine (SVM) is a method to classify the linear and nonlinear data. To transform the original training data into higher dimension, a nonlinear mapping is used. The features of global optimization and high generalization ability are the benefits of SVM. Like ANN, when comparing with tradition approaches, SVM removes over fitting issues and provides a sparse solution.

3. Data Mining Tools

3.1 Orange

Orange is the perfect software set for machine learning and data discovery. It is an excellent data visualization and software-based part. It was written in the Python computer language. It is a software-based component and the orange components are called “widgets”. These widgets range from data visualization and pre-planning to assessment of algorithms and predictions [6].

3.2 Weka

Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best for data analysis and forecasting. It includes algorithms and visualization tools that support machine learning. Weka has a GUI that can easily access all its features. It was written in JAVA programming language. Weka is a data search and development company. Editing Seeing Supports key search and retrieval tasks. It works on the assumption that data is available in a folder format. Weka can access SQL databases through database connections. Repeat data / results returned from an inquiry.

3.3 Rapid Miner

Rapid Miner is a freeware search tool. It can edit information, Used for machine learning and propagation. It provides a variety of products for setting up new data discovery processes and prediction setup analysis.

3.4 KNIME

KNIME is an open source software to create scientific application applications and services. This information search tool helps you to understand information and to design data science traffic. KNIME is an integrated platform for data analysis and reporting. Developed by KNIME.com AG. The situation is working on the concept of a modular data pipeline. Combined with machine learning and data discovery components, KNIME combines. KNIME is widely used for medical research. In addition, it can analyze customer data, analyze and analyze customer data. Works well for financial data analysis and business intelligence. KNIME has brilliant features such as rapid scalability and scalability. Users are familiar with KNIME in less time, and it also provides access to users who use predictive analytics. KNIME uses the installation of nodes to pre-arrange data for analysis and transparency.

3.5 Sisense

Users are familiar with KNIME for a limited time and also allow users to use predictions. KNIME uses the installation of nodes to plan data for analysis and transparency.

3.6 Mango DB

It's free, various platforms It is a database management system that is not related to open source documentation. It is under the NO_SQL database category.

3.7 Python

Finding It Analysis and educational data are used for estimation. Puthon Numpy and other programs are used to accelerate the exercise and to build relationships with other packages that support interoperability with other packages in the Python ecosystem.

4. Data Mining Process

There are many processes in this model. They are data preparation (selected from storage database), data selection and transformation (prepared data is selected and transformed) and evaluate using data mining technique. Figure 1 shows the process of the data mining model [7].

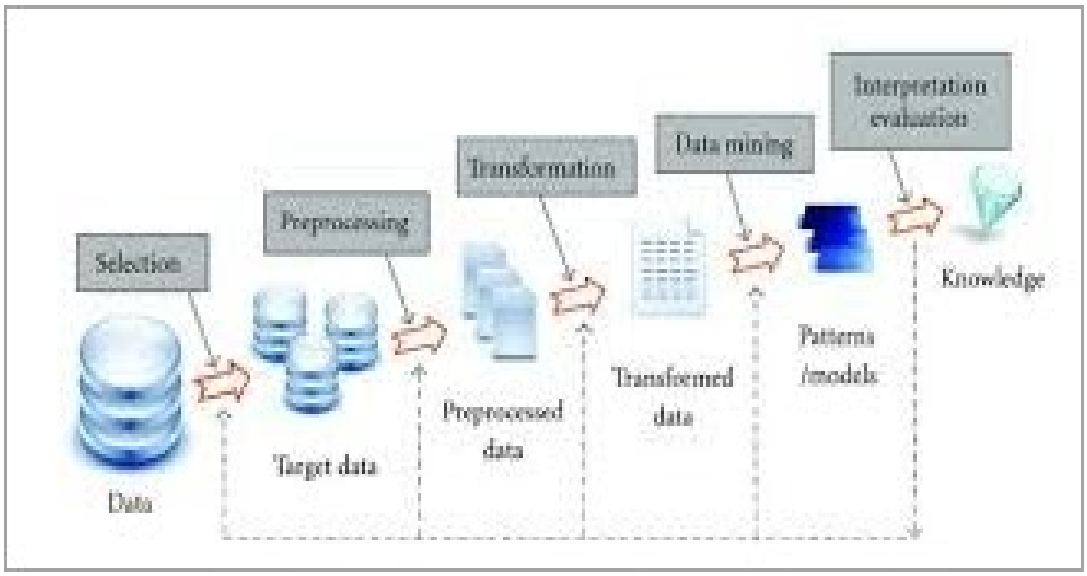

Figure 1. Data mining process model.

4.1 Student Assessment Dataset Preparation

The dataset are obtained from students' performance at Third Year in the university. In students' performance includes attendance, assignment, quiz, project, lab test (practical exam) and exam (paper exam).

4.2 Data Selection and Transformation

Table 1. Data Description and Possible Values

Variable	Description	Possible Value
ATT	Attendance in each semester	{Poor, Average, Good}
ASS	Assignment	{Yes, No}
Q	Quiz	{Yes, No}
P	Project	{Poor, Average, Good}
L	Lab Test	{Yes, No}
EM	Exam Mark	{A, B, C, D, E}

- ATT- Attendance of Student. At least 75% attendance in one semester examination. ATT is classified into three parts: Poor- <60%, Average - >60% and Good - > 80%.
- ASS - This is assignment performance. ASS is classified two parts: Yes (Assignment submitted) and No (not

submitted).

- Q - This is the general studies (answering the quiz) on one subject. This is also Yes (submitted) and No (not submitted).
- P - Team work performance (Project). This is classified into three: Poor, Average and Good level.
- L - This is the work in Lab room. This is two parts: Yes (complete lab work) and No (not complete lab work).
- EM - This is the three hours testing and classified five parts: A 80-100, B 61-80, C 41-60, D 21-40, E 0-20.

4.3 Applying Data Mining Technique (Naïve Bayes)

Steps of Naïve Bayes Algorithm:

- Step 1: Scan the dataset
- Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]
- Step 3: Apply the formulae

Where:

- n = the number of training examples for which $v=v_j$
- n = the number of examples for which $v = v_j$
- $p\left(\frac{ai}{vj}\right) = \frac{n\_c + m * p}{n + m}$
- m= the equivalent sample size [number of attributes]

- Step 4: Multiply the probabilities by p
- Step 5: Compare the values and specify the attribute values to one of the predefined set of class.

5. Results and Discussion

The total number of students is 96. Among them, 50 students are examined during each semester in 2018-2019 academic year in university. 50 students dataset are described below.

Table 2. Students' Performance Dataset

No	ATT	ASS	Q	P	L	EM
1	Good	Yes	Yes	Good	Yes	A
2	Good	Yes	No	Average	Yes	A
3	Average	No	No	Average	No	A
4	Good	No	No	Good	Yes	A
5	Good	No	Yes	Average	Yes	A
6	Average	No	No	Average	Yes	A
7	Poor	No	No	Average	Yes	B
8	Average	Yes	Yes	Poor	No	A
9	Poor	No	No	Poor	No	C
10	Good	Yes	Yes	Average	No	A
11	Good	Yes	Yes	Good	Yes	A
12	Good	Yes	Yes	Average	Yes	A
13	Good	Yes	No	Average	No	A
14	Good	Yes	Yes	Good	No	A
15	Average	Yes	Yes	Average	Yes	A
16	Poor	Yes	Yes	Average	Yes	B
17	Good	Yes	Yes	Average	Yes	B
18	Poor	Yes	Yes	Average	Yes	B
19	Good	No	Yes	Average	Yes	B
20	Average	Yes	No	Poor	Yes	B

21	Poor	No	Yes	Average	No	C
22	Average	Yes	Yes	Poor	Yes	C
23	Average	No	No	Poor	Yes	C
24	Good	Yes	Yes	Poor	Yes	B
25	Poor	Yes	Yes	Poor	Yes	C
26	Poor	No	No	Poor	Yes	D
27	Good	Yes	Yes	Good	Yes	A
28	Good	Yes	Yes	Good	Yes	B
29	Good	Yes	Yes	Average	Yes	B
30	Average	Yes	Yes	Good	Yes	B
31	Good	No	No	Good	Yes	B
32	Good	Yes	Yes	Average	Yes	B
33	Average	No	Yes	Average	Yes	C
34	Good	No	No	Good	Yes	C
35	Average	No	Yes	Average	Yes	C
36	Average	No	No	Poor	Yes	C
37	Average	Yes	No	Average	Yes	C
38	Poor	No	Yes	Average	Yes	D
39	Poor	No	Yes	Average	Yes	C
40	Good	No	No	Poor	No	C
41	Poor	No	Yes	Poor	Yes	D
42	Poor	No	No	Poor	No	D
43	Good	Yes	Yes	Good	Yes	B
44	Average	Yes	Yes	Good	Yes	B
45	Average	Yes	Yes	Good	Yes	C
46	Average	Yes	Yes	Poor	No	D
47	Poor	No	Yes	Poor	Yes	D
48	Poor	No	No	Poor	Yes	D
49	Good	Yes	Yes	Average	Yes	B
50	Poor	No	No	Good	No	D

6. Conclusion

Data mining methods can be used to make effective decisions and to improve the analysis of students' performance in one semester. Naïve Bayes classifier evaluates the assessment of students by using 50 students, six attributes (Attendance, Assignment, Quiz, Project, Lab Exam, and Exam Marks) and predicts the student performance at the end of each semester. The result makes to realize students who essential special responsiveness to which parts and teachers that students are less interesting in which performance and that they should be encouraged in which. Finally, the result shows that the accuracy in educational students' performance by using Naïve Bayes algorithm is over 90 % very high.

7. Acknowledgements

We would like to acknowledge all of our colleagues at University of Computer Studies, Hinthada for their care and help through our research effort. Big thanks a lot to our families for their help and support.

References

[1] Hafez Mousa, Ashraf Maghari. (2017). "School Students' Performance Prediction Using Data Mining Classification", pp. 136-141, JARCCE, 2017.

[2] A. Dinesh Kumar, R. Pandi Selvam, V. Palanisamy. (2019). "Prediction of Student Performance using Hybrid Classification", IJRTE, pp 6566-6570, 2019.

[3] Y. Divyabharathi, P. Someswari. (2018). "A Framework for Student Academic Performance Using Naïve Bayes Classification", JAET, pp. 1-4, 2018.

[4] Brijesh Kumar Baradwaj. (2011). "Mining Educational Data to Analyze Students' Performance", pp. 63-69, 2011.

[5] Ahmed S. J. Abu Hammad. (2018). "Mining Educational Data to Analyze Students' Performance (A Case with University of Science and Technology Students)". Pp. 56-65, 2018.

[6] Pooja Thakar, Anil Mehta, Ph. D, Manisha, Ph.D. (2015). "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue". Pp. 60-68, 2015.

[7] Kalpesh P. Chaudhari, Riya A. Sharma, Shreya S. Jha, Rajeshwari J. Bari. (2017). "Student Performance Prediction System using Data Mining Approach", IAJRCCE, pp. 833-839, 2017.