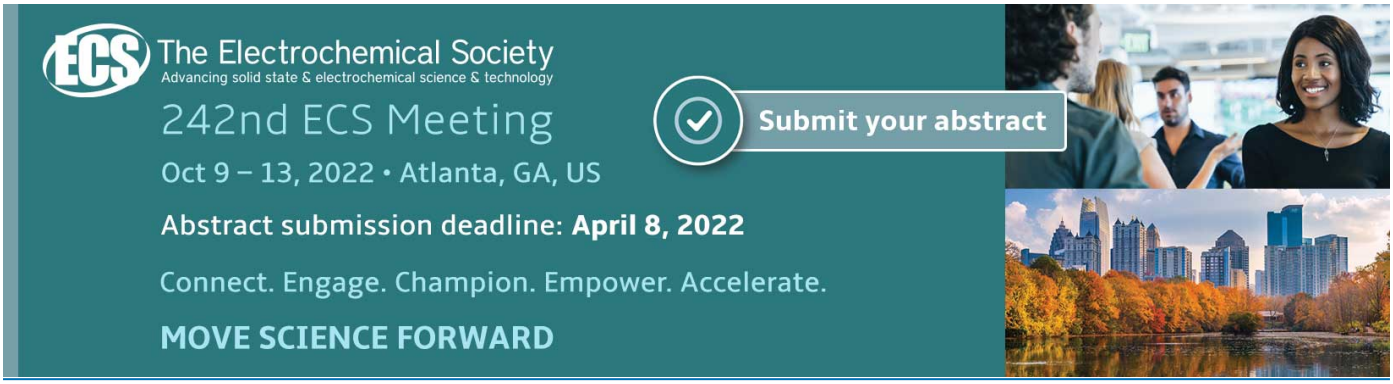
Data Mining of Family, School, and Society Environments Influences to Student Performance



IOP Conf. Series: Materials Science and Engineering **420** (2018) 012090 doi:10.1088/1757-899X/420/1/012090

Data Mining of Family, School, and Society Environments Influences to Student Performance

Yo Ceng Giap[1]*, Nico Leonardi[1], Bayu Waseso[2] and Robbi Rahim[3]

[1]Department of Information Technology, Universitas Buddhi Dharma, Tangerang, Indonesia
[2]Faculty of Computer Science, Universitas Mercubuana, Jakarta, Indonesia
[3]School of Computer and Communication Engineering, Universiti Malaysia Perlis, Kubang Gajah, Malaysia

*cenggiap@buddhidharma.ac.id

Abstract. It is a common fact that every parents wants their children to get achievements, especially academic achievements. Demands by demands then addressed to their children to get higher and higher score in every subjects. However, parents and teachers are often forgot about one thing, besides of just demanding the children to get high scores, they also responsible in providing the best conditions to boost the children's potential and performance. To help solving this issue, the author used data mining model in binary classification approach. Data mining methods that author chose was decision tree due to its high accuracy comparing to the neural network and support vector machine that have lower accuracy than decision tree. Despite of previous result that is the main factor that affects student performance, in this research, we also found several factors of environments that may affects the student performance

1. Introduction

Academic achievement is a part of all students desires, especially their parents[1][2]. Parents would do anything to make their children can achieve more in their school, such as apply the children to a course class, calling private tutor, etc[3]. But, besides of all those classes, there are things that are being forgotten by parent to boost the children potential. It is the influences from the children surroundings. In this research, the author put mathematic subject student results as a sample, because there is a stereotype that students who good at math, is clever especially in Indonesia. We hope that this research existence can make parents have more cooperation and communication with teachers to give the best conditions for their children's potential, the number of students that are stressed will decrease, and the parents and teachers can give more attention to the children.

This research used data mining methods in the process and with the significant development of information technology, databases are being used more often nowadays[4]. Those lines of data can be processed to make an important information that can be a crucial clue for developing things[5][6]. To process those data, we can use data mining model. Data mining model can find patterns from many lines of data, which that pattern itself contains informations[7]–[10]. All data mining processes in this research will be done by using a tool named Rapid Miner[11], [12].

2. Methodology

This research is based on use data warehouse and data mining to predict student academic performance in schools: a case study (perspective application and benefits)[13] article with several adjustments of conditions with the conditions in Indonesia especially scoring system. There are 2 data sources in this research, first is secondary data which the author got from journal attachment and primary data which the author got from several students in Jakarta and Tangerang city. Total data that are being used is 425 data, consists of 395 secondary data from Gabriel Pereira school and Moushinho da Silveira school and 30 primary data from a social group named Dharma Subha Temple, because many members of that group are from several schools from various regions such as Poris Indah School, Mutiara Bangsa School, etc. The attributes and variables that are being examined are based on the source journal which the author will explain on table 1.

Table 1. Attributes and Variables

A & V	Description
Sex	Student sex (binary) "F"(*female*), "M"(*male*).
Age	Student age (numeric) 15 until 22.
Address	Student's house location (binary) "U" (*urban*), "R"(*rural*).
Pstatus	Parents status (binary) T (*together*), A(*apart*).
Medu	Mother's education * (numeric) no education is 0, primary graduated is 1, Junior high graduated is 2, Senior high graduated is 3, Bachelor or diploma graduated is 4.
Mjob	Mother's job # (polynominal) education equals "*teacher*", medical equals "*health*", civil services (police, etc) equals "*services*", just at home equals "*at_home*", other equals "*other*".
Fedu	Father's education * (numeric no education is 0, primary graduated is 1, Junior high graduated is 2, Senior high graduated is 3, Bachelor or diploma graduated is 4.
Fjob	Father's job # (polynominal) education equals "*teacher*", medical equals "*health*", civil services (police, etc) equals "*services*", just at home equals "*at_home*", other equals "*other*".
Guardian	Main guardian of student (polynominal) "*father*", "*mother*", "*other*".
Famsize	Family size (binary) "GT3"(*greater than 3*), "LE3" (*less equal 3*).
Famrel	Family relationship (numeric) 1 until 5.
Reason	Reason to choose current school (polynominal) near with house equals "*home*", school reputation equals "*reputation*", *course preference* equals "*course*", "*other*".
Traveltime	Travel time from home to school (numeric) less than 15 minutes equals 1, 15-30 minutes equals 2, 30-60 minutes equals 3, more than 1 hour equals 4.
Studytime	Study time outside the school (numeric) >2 hours equals 1, 2-5 hours equals 2, 5-10 hours equals 3, >10 hours equals 4.
Failures	Failures at math subject exams in a year (numeric) 0 until 4
Schoolsup	School supplementary lessons (binary) "*no*", "*yes*".
Famsup	Family supplementary lessons (binary) "*yes*", "*no*".
Activities	Extracurricular activites joined (binary) "*yes*", "*no*".
Paidclass	Join any math course (binary) "*yes*", "*no*".
Internet	Any internet access at home (binary) "*yes*", "*no*".
Nursery	Joined nursery school in the past (binary) "*yes*", "*no*".
Higher	Students want to study in higher level (binary) "*yes*", "*no*".
Romantic	Student is in relationship (binary) "*yes*", "*no*".
Freetime	Freetime outside the school (numeric) 1 until 5. 1 rarely, 5 very often
Goout	Student going out with friends (numeric) 1 until 5. 1 rarely, 5 very often
Walc	Alcohol consumption on weekend (numeric) 1 until 5. 1 few, 5 very much

Dalc	Alcohol consumption on normal days (numeric) 1 until 5. 1 few, 5 very much
Health	Student's present health condition (numeric) 1 until 5. 1 terrible sick, 5 very healthy
Absences	Absences taken in a year (numeric) 0-93
G1	First year math final exam score. (numeric) 0,5,10,15,...,95,100
G2	second year math final exam score. (numeric) 0,5,10,15,...,95,100
G3	National exam math final exam score. (numeric) 0,5,10,15,...,95,100
Grade	An additional attribute about G3 (binary), G3>69 equals "Pass", G3<70 equals "Fail". Because in Indonesia 70 is the most common least score to pass a math exam.

Rapid Miner is a program licensed under the GNU Affero General Public License version 3[14]. This program has a convenient interface, where analysis is configured in process view. Rapid Miner uses a modular concept in the process, in which each analysis step is illustrated by an operator in the analysis process. These operators have ports for inputs and outputs where they can communicate with other operators for the purpose of receiving data input or passing the modified data and generating models through the included operator.

Thus, a process that has been created can be reused immediately for similar problems, a model that has been generated once it is loaded and applied or the results obtained can be viewed solely for a method that can provide the most promising research success. The results can be shifted and placed in processes where the results will be reprocessed with a special operator and prepared for further processing.

In addition, for local repositories, which are stored in file systems on computers, there are Rapid Analytics agencies that create wider functions, allowing easier analysis execution. Users can not only save the process, but also execute it with the Rapid Analytics agency. This means that the analysis is fully implemented in the background and users can continue to work at the same time without worrying about slow CPU performance and intensive memory computation

3. Result and Discussion

The following is the result of the decision tree generated by the Rapid Miner tool:

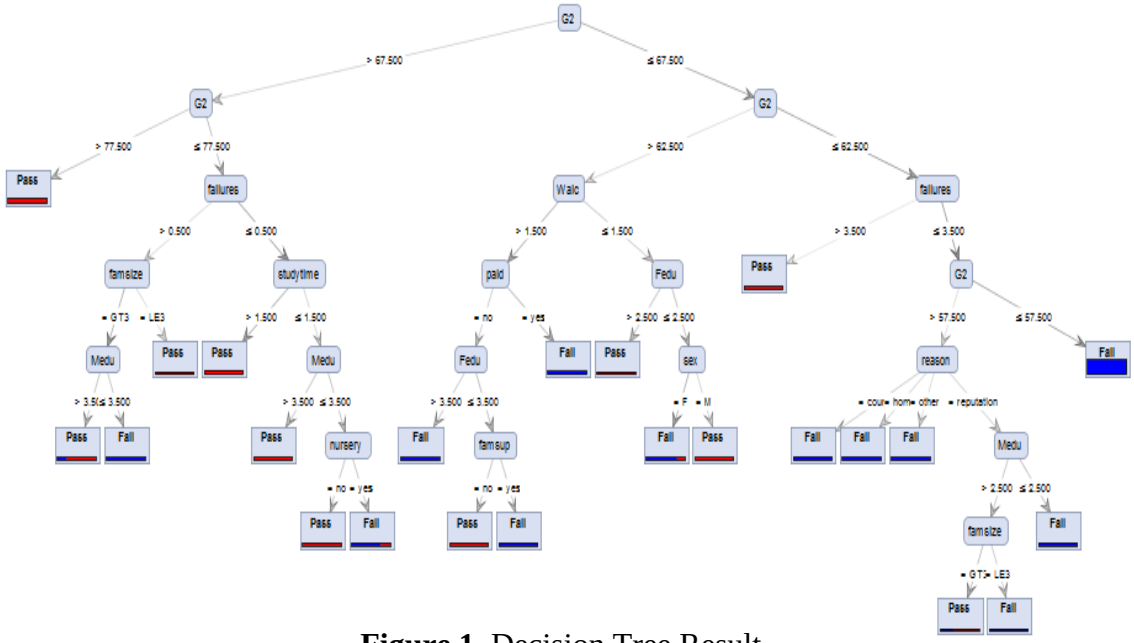


Figure 1. Decision Tree Result

Based on decision tree results obtained from the Rapid Miner tool process, it can be concluded that G2 (2nd year score) is the most important variable in graduation and achievement of a student. It has been proven with the G2 variable that decides on the first and second level nodes. If the value of G2 is greater than 67.5, then in next branch it will be checked again, if G2 is greater than 77.5, then certainly the student will pass on the next test. If not, then the next check is whether failures (experiencing failure) is greater than 0.5 or not. In other words, whether the student has experienced failure or never experienced it, because the input received is only in integers 0.1,2,3,4.

If the student has experienced a failure, then the next attribute, famsize (number of family members) will be examined. If famsize is a GT3 (Greater than 3 / more than 3), then Medu will be checked next (Mother's last education) if it is greater than 3.5 or in other words is 4 which means passing college, then the student is predicted to pass. If the mother's last education is below college graduation, then the student is predicted to fail or fail on the next test. Next, back to 2 node level above, if the famsize is LE3 (Less equal 3 / below or equal to 3 people) then the student is predicted to pass.

Back to third level node from the top, if the student has never experienced a failure at all, then study time will be checked (out-of-school time). If the study time attribute of the student is more than 1.5 (at least studying with a minimum duration of 2 hours), then the student is predicted to graduate. However, if less than 1.5, then check the value of Medu, if Medu value is above 3.5 (at least must pass college), then the student is predicted to pass. If the Medu value is below or equal to 3.5, then the next attribute that will be checked is nursery (never entered nursery). If the value is a yes, then the student is predicted to fail, and if the value is a no, then the student is predicted to pass.

Then return to the top level node, if the value of G2 is smaller than or equal to 67.5, then G2 will be re-checked, whether G2 is greater than 62.5. If G2 is greater than 62.5, then the next attribute to be examined is Walc (the level of alcohol consumption at the weekend), whether it exceeds the level of 1.5 or not. If it exceeds the level 1.5, then the next attribute we check is paid (does the child attend a paid course outside school?).

If the child is attending a paid course class, the child is predicted to fail, but if not then the Fedu attribute (Father's last education) will be examined. If Fedu has value more than 3.5 (at least graduated from college), then the child is predicted to fail, if the Fedu value is smaller or equal to 3.5, then check the famsup attribute (additional education by parent), if there is additional education from the parents, This child is predicted to fail, but otherwise, if this student does not get additional education from parents at home, then this child is predicted to pass.

Return to the third level of node from the top, when Walc's attribute is checked. If the weekly alcohol consumption rate is below or equal to 1.5, then the Fedu attribute will be checked. If the value of Fedu is greater than 2.5 (at least have to pass high school or equivalent), then the student is predicted to pass. If the value of Fedu is below or equal to 2.5, then the sex attribute (sex) will be checked next. If the student is a male, then he is predicted to graduate, but if the student is female, then the student is predicted to fail. Next, go back to the 2nd level of node from the top where G2 is checked whether it is more than 62.5 or not. If the value of G2 is below or equal to 62.5, then the failures attribute will be checked. If the failures value is greater than 3.5, then the child is predicted to pass. If the failures value is below or equal to 3.5, then the attribute G2 will be re-checked, whether the value of G2 is greater than 57.5 or not. If the value of G2 is smaller or equal to 57.5, then the child is

predicted to fail. But if the value of G2 is greater than 57.5, then the next attribute that will be checked is reason (the reason for choosing the current school).

If the reason for choosing a school is because of course preference, close to home, or choose the "other" option then the child is predicted to fail. However if the reasons for choosing a school is because of the school reputation, then we will examine Medu attributes. If the value of Medu is smaller or equal to 2.5 (at least graduated from high school equivalent), then the child is predicted to fail. If the Medu value is greater than 2.5, then at the last part we will check famsize (number of family members). If the number of family members is smaller or equal to 3 people, then the child is predicted to fail. But, if the number of family members is greater than 3 people, then the child is predicted to pass.

4. Conclusion

From the total 33 attributes and 425 rows of reference data used at the beginning of the study, it turns out that only 12 attributes really affect the outcome or achievement or graduation from upper secondary students. Those 12 attributes are sex, Medu, Fedu, famsize, reason, studytime, failures, famsup, paid, nursery, walc, and g2. 2nd year exam score (G2) is the most influential attribute compared to all other attributes. Both parents' education also has an important role to play in their children's achievement in school. Then failures (number of failures) also have an important effect. Learning from failures proves to play an important role in a child's achievement. Alcohol consumption also has a great effect on achievement. For other attributes, such as famsup, paid, nursery, sex, famsize, little effect on student achievement because of its position in the tree that are in the bottom of the root.

The author retrieves the necessary data to conduct the research, the data taken is secondary data from the machine learning repository uci website with the title of student performance data set for the development of science by Paulo Cortez as early journal journalist and collector of data in two schools in the country of Portugal. From these things, the authors get the data set as the basis of research, which is 395 rows of high school student data. The author also did dig the primary data in the form of questionnaires made online using Google Form. Collect online that allows to perform legally accountable ie Dharma Subha Vihara located at Poris Indah Jl. Pinus Raya Blok E no. 489, Cipondoh, Tangerang, 15148 and got 30 respondents.

From the comparison of the three methods used can be concluded that the accuracy of Neural Network is superior to the accuracy of the Support Vector Machine method. While the accuracy of the Decision Tree method can look more superior than the method of NN and SVM. And the accuracy of the data increases as the amount of data increases

References

[1] Adiyarta K Napitupulu D Rahim R Abdullah D and Setiawan M, Apr. 2018 Analysis of e-learning implementation readiness based on integrated elr model *J. Phys. Conf. Ser.* **1007**, 1 p. 012041.

[2] Napitupulu D *et al.*, Jan. 2018 Analysis of Student Satisfaction Toward Quality of Service Facility *J. Phys. Conf. Ser.* **954**, 1 p. 012019.

[3] Rangka I B, 2018, Datasets: Student Learning Performance Based on Geometry Test. Open Science Framework.

[4] Siregar D Arisandi D Usman A Irwan D and Rahim R, Dec. 2017 Research of Simple Multi-Attribute Rating Technique for Decision Support *J. Phys. Conf. Ser.* **930**, 1 p. 012015.

[5] Harliana P and Rahim R, Dec. 2017 Comparative Analysis of Membership Function on Mamdani Fuzzy Inference System for Decision Making *J. Phys. Conf. Ser.* **930**, 1 p. 012029.
[6] Ahmar A S *et al.*, 2018 Modeling Data Containing Outliers using ARIMA Additive Outlier (ARIMA-AO) *J. Phys. Conf. Ser.* **954**, 1.
[7] Rahim R *et al.*, 2018 Block Architecture Problem with Depth First Search Solution and Its Application *J. Phys. Conf. Ser.* **954**, 1 p. 012006.
[8] Rahim R Ahmar A S Ardyanti A P and Nofriansyah D, Dec. 2017 Visual Approach of Searching Process using Boyer-Moore Algorithm *J. Phys. Conf. Ser.* **930**, 1 p. 012001.
[9] Rahim R Nurarif S Ramadhan M Aisyah S and Purba W, Dec. 2017 Comparison Searching Process of Linear, Binary and Interpolation Algorithm *J. Phys. Conf. Ser.* **930**, 1 p. 012007.
[10] Rahim R Zulkarnain I and Jaya H, 2017 A review: search visualization with Knuth Morris Pratt algorithm in *IOP Conference Series: Materials Science and Engineering* **237**, 1 p. 012026.
[11] Sreenivasa Rao K Swapna N and Praveen Kumar P, 2018 Educational data mining for student placement prediction using machine learning algorithms **7**, 1.2 p. 43–46.
[12] Zheng Y U, 2015 Trajectory Data Mining : An Overview *ACM Trans. Intell. Syst. Technol.* **6**, 3 p. 1–41.
[13] Kurniawan Y and Halim E, 2013 Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application and benefits) in *Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)* p. 98–103.
[14] Hartama D Mawengkang H Zarlis M and Rahim R, 2017 A Research Framework of Disaster Traffic Management to Smart City in *2017 Second International Conference on Informatics and Computing (ICIC)* p. 1–5.