International Journal of Data Science and Analysis
2020; 6(5): 120-129
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200605.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Factors Influencing Secondary School Student's Performance Through Variable Decision Tree Data Mining Technique

Yousaf Ali Khan[1, 2]

[1]School of Statistics, Jiangxi University of Finance and Economics, Nanchang, China

[2]Department of Mathematics and Statistics, Hazara University Mansehra, Mansehra, Pakistan

Email address:
yousaf_hu@yahoo.com

To cite this article:
Yousaf Ali Khan. Factors Influencing Secondary School Student's Performance Through Variable Decision Tree Data Mining Technique. *International Journal of Data Science and Analysis.* Vol. 6, No. 5, 2020, pp. 120-129. doi: 10.11648/j.ijdsa.20200605.11

Received: January 17, 2020; **Accepted**: September 10, 2020; **Published**: September 25, 2020

Abstract: Schools are considered as the backbone for long-term economic progress. No country can develop without increasing their education level. Despite the fact that the Portuguese population shows a brilliant development in their educational level from last decade, but still Portugal lies on the tail surrender of Europe in statistics because of excessive levels of student failure. Primarily, this costs a lot better in the middle of the elegance of Mathematics and Portuguese. On the other hand, the field of data mining (DM), the purpose of extracting the high-stage knowledge of raw statistics, automatic gear compelling offer to a useful source of training domain. This paper pursues to improve the overall performance of middle school students of Portugal through two variables decision tree, which is a favorable approach to data mining used for classification, prediction and factors explored with the help of their significance. Results shows that, provided the first and / or second interval school grades, awesome prediction accuracy can be achieved. Despite the success of students strongly influenced by father's job assistance; evaluation has clearly shown that there are also other elements (such as learning time, mother's occupation, the desire of higher education, the paid-classes and the travel time from home and school, etc.) are important elements which have great impact on the performance of students in secondary school education in Portugal. As a direct result of this study, through which specialize in these factors and create a kind of policy is mainly based on studies in the country width exceptional level of education may increase at the secondary level that produces goose bumps to the stage of higher education in Europe.

Keywords: Data Mining in Education, Secondary School, Decision Tree, Performance, Classification, Europe's

1. Introduction

Schooling is a key issue to achieve a protracted time period of financial development. Over the past decade, the Portuguese academic level has stepped forward. But, record keeping Portugal in Europe tail stopped because of high student failure and defeat quotes. For example, in 2006 the early school leaving rates in Portugal is 40% for 18 to 24 years of age, even as the union average European value is only 15% [1-2]. Especially, a failure in the core classes of Mathematics and Portuguese (local language) was of tremendous importance, due to the fact they offer essential skills for success in the closing lecture topics (e.g. physics or history, and so on.).

Alternatively, the interest in enterprise intelligence (BI) / record mining (DM) [3], appears as progress the fact generation, major for exponential explosion of commercial enterprise and database organization. All these facts hold valuable records, along with trends and styles, which can be used to improve decision-choice and optimize achievement. But, professional human control and can overlook important details. As a result, the opportunity is to use automated equipment to study the raw information and extract facts excessive rate of interest for decision makers. Arena education provides fertile ground for business intelligence program, due to the fact that there is some asset information (e.g., a conventional database, on-line web pages) and a variety of hobby organizations (e.g., students, instructors, directors or alumnus) [4]. For example, there are some interesting questions for this area that will be answered using

techniques BI / DM [5]: which are students take most credit hours? which is likely to return for more classes? what form of guidance can be offered to attract more students? what the main motive for transfer students? it is reasonable to expect the overall performance of students? what are the factors that affect student achievement? This paper will be awareness in the two closing questions. Type and predict the overall performance of students is an important device for every educator and student, arguing that it can help the higher the know-how of this phenomenon and the long-term fix. For example, professional school may want to undertake remedial measures for students who are weak (e.g. remedial classes).

In effect, few studies have addressed the subject of comparable. [6] Applied DM techniques are based entirely in association policy that allows you to choose a tertiary college students vulnerable from Singapore for remedial classes. Enter a closed variable demographic attributes (e.g., gender, area) and the performance of lecturers during the last years and the answer proposed allocation outperform conventional manner. In 2003 [6], on-line student scores on the Michigan Royal Colleges have been modeled using three classification techniques (i.e. binary: pass / fail; three degrees: low, medium, redundant, and 9 levels: from 1 - the lowest grade up 9 - the maximum score). Database covered 227 samples with feature line (e.g., a variety of solutions to be corrected or attempt to homework) and the consequence was obtained with the help of an ensemble classifier (e.g., decision trees and neural networks) with the cost of accuracy of 94% (binary), seventy-two% (three grade) and 62% (nine classes). Kotsiantis et al. (2004) conducted various DM algorithms to predict the performance of computer science technological know-how of students of a university program within mastering [7]. For each student, some demographics (e.g. sexual intercourse, age, marital status) and performance attributes (e.g. marking tasks given) has been used as the input of a skip binary / fail classifier. Satisfactory answers be obtained using Naive Bayes method with an accuracy of seventy-four%. Also, be observed that the external value of the college has the effect of a mile higher than demographic variables. Additionally, Pardos et al. (2006) notes collected from the gadget les internet about the united states of America in eighth grade mathematics assessment [8]. The authors adopt the regression method, where the goal becomes to predict ranked checks mathematics based on the talent. The author uses a Bayesian network and the end result is satisfactory into the prediction error of 15%. In this research, I apply the methods developed my own two decision trees variables and analyzed the facts of students from two high schools Portuguese who had previously been analyzed and advocates that there is a desire from the teeth prediction undergraduate efficient to be developed that correctly classify the facts and hope more as it should be for details see [9]. The goal is to create a category of efficient information helpful in predicting student performance overall use of two classes of the center (i.e. Mathematics and Portuguese) and to identify the key elements that have an effect on the performance of secondary school students is that facilitate the selection made via the tree selection approach is the fact mining variable.

The rest of the paper was prepared as follows. Section 2 provides, materials and methods Section 3 provides variable decision tree construction Section 4 offers student's overall performance, results and discussion Finally, section 5 concludes this research. All technical details are given in the Appendix.

2. Materials and Methods

2.1. Student Data

In Portugal, the secondary training consists of three years of training, the previous nine years of basic education and training is accompanied by a higher use. Maximum of students become part of the training device and loose general public. There are many guides (e.g., science and technology, art visible) which shares the core topics that include Portuguese and Math's. Unlike some countries (e.g. France or Venezuela), a Twenty-factor levels used scale, where zero is the lower class and twenty is a perfect score. At some point in the school year, students are evaluated in three periods and the final evaluation (G3 Table 3) in accordance with this last class we see will not forget statistics collected throughout the 2005-2006 school from the school community, from where alentemol of Portugal. Although there is a tendency for investment growth generation statistics from the government, the general public of the Portuguese public school data systems are very poor, relying largely on a sheet of paper (which became the contemporary case). Therefore, the database turns into built assets: review of the school, based on pieces of paper and together with some attributes (i.e. three values of the period and the number of school absences); and questionnaires, used to supplement information previously. The questionnaire was mainly based on closed questions (i.e. with a choice of pre-set) related to demographics (e.g., employment, family income, mother job) a lot, (e.g. alcohol intake) social / emotional and school-related (e.g., a variety of failures grandeur of past ago) variable that has been anticipated to have an effect on the overall performance of students. Questionnaires by academic experts and checked on a small set of 15 students to be able to get feedback. The last model contained 37 questions in a single A4 sheet and answered in splendor through the 788 students. Person, 111 solutions have been discarded due to lack of identity information (important for merging the school review). Furthermore, the statistics become integrated into a dataset associated with Mathematics (with 395 examples) and Portuguese (649 facts) lesson. At some stage in the preprocessing level, several functions have been discarded because of the lack of value discriminatory. For example, some respondents spoke again about the advantages of their family (most likely because of privacy issues), at the same time almost one hundred% of students live with their parents and have a personal computer at home. Final attribute evident in Table 3 Appendix A2, where the final four rows indicate the

variables are taken from reviews of the school [9].

2.2. Two Variable Decision Tree Data Mining Technique

The fact mining is the extraction of implicit information, previously unknown and rotationally useful from the record. In addition, it was miles' extraction from large databases into useful information or statistics and information called comprehension. Mining information continues to be included in the strategy to discover and describe the structural pattern in the information as a device to help that statistical and predictive makeup. Statistics mining consists of five essential elements. First, extract, the facts and transactions cargo into the device data warehouse. Second, maintain and manage the facts in a multidimensional database gadget. Third, offer a record gain entry to commercial companies and professional analysts. Fourth, statistical analyzes with the aid of a software program application. Fifth, presenting the data in a useful layout, along with a chart or table. Many facts mining techniques are closely associated with several machine learning. Others related to the strategies that have been developed in fact, now and again called exploratory statistical evaluation [10].

In a note and get to know the machine, the category is difficult to identify which of the hard and fast on the new speech class belongs, on the idea of a set school records that contain comments that class membership understood. Among the many different strategy decision tree class is one of the most famous gadgets powerful modeling study nonparametric approach is used for both prediction and classification problems. A decision tree is a tree-dependent classifier that do break until we look at the internal node and predicts the target sample in a knot grandeur leaves. With the simplicity and transparency of them, a probability tree is widely used in mining records [11-12]. Novel two variable selection tree based entirely on statistical coefficient maximum, which is a useful statistical approach of mining with the extra ability to explore and rank the elements in step with their interests. Two variables tree selection works in principle reliance size and classify the data into the company; maximum use of statistical coefficient (MIC) as an index of categories, where the association is quite one-of-a-kind in the two groups of data sets. Maximal information coefficient (MIC) is primarily based on the facts of each proposed recently by using Rashef et al. (2011) as statistical degree of dependency to seek a new affiliate in the set of information regardless of the shape of its objectives. At the heart of this definition is the fact each naive to predict I_{MIC} [X; Y] are calculated using a binning scheme based information. Let n_x and n_y, respectively, showing the amount of bins that is imposed on *x* and *y*-axis. MIC binning scheme is chosen so that (i) the total amount of bins $n_x n_y$ no longer exceed some specific consumer value of *B* and (ii) the value of the ratio;

$$MIC\{x;y\} = \frac{I_{MIC}\{x;y\}}{Z_{MIC}} \quad (1)$$

Where $Z_{MIC}=log_2(min\ (n_x, ny))$, is maximized.

The ratio is provided in Eqn. 1 is calculated using a binning scheme based on this record, is how the MIC defined. Note that, because I_{MIC} bounded above by way Z_{MIC}, the values of MIC will continue to fall between 0 and 1. Therefore, the MIC is a measure of novel associations for large data sets. see [13-17] for a full assessment and estimated maximal statistical coefficient for the two variables.

In this study, we will build a new two-variable decision tree for secondary school facts from Portugal use somewhat correlated variables; Math and Portuguese value and reach of factors that impact on the overall performance of students in secondary schools. Which allows to predict students' overall performance in addition to coverage reasons they make to improve the performance of high school students.

3. Construction of Decision Tree for Student Performance

We start from the planned two period values (e.g. Mathematics and Portuguese) of secondary school students. As proven in Figure 1. Two variables are highly correlated and we expect one of the alternatives because there might be an effective high correlation between them. Perhaps there are thirty attributes each with extraordinary for middle school students who have direct effect on performance. Details related student attributes and explanation was given in Table 3. Appendix A2.

We found all the attributes one by one and found the size dependence (MIC) for each level attribute as shown in Table 2. Appendix A1. We classify all the attributes in institutions; variation in their degrees and vice versa having significant difference at their levels. Among thirty, eight had significant differences in the level of their MIC.

As evidenced in Table 2. Appendix A1. Dominant element has both the size difference in their ranges are provided in the Table 1 below.

Table 1. *Factor selection for classification at chide node of decision tree.*

Factors	MIC of Factors Levels	Difference of MIC between levels	Classification Factor
Study Time	0.5795869; 0.8419889	-0.26240197	
Mother Jo	0.5507668; 0.8093887	-0.25862190	
Paid Class Higher	0.5021382; 0.6301857	-0.12804757	
Education	0.5992320; 0.4509993	0.14823272	Father Job
Father Job	0.8450145; 0.5777136	0.267300936	
Family Relation	0.5885900; 0.7958309	-0.20724086	

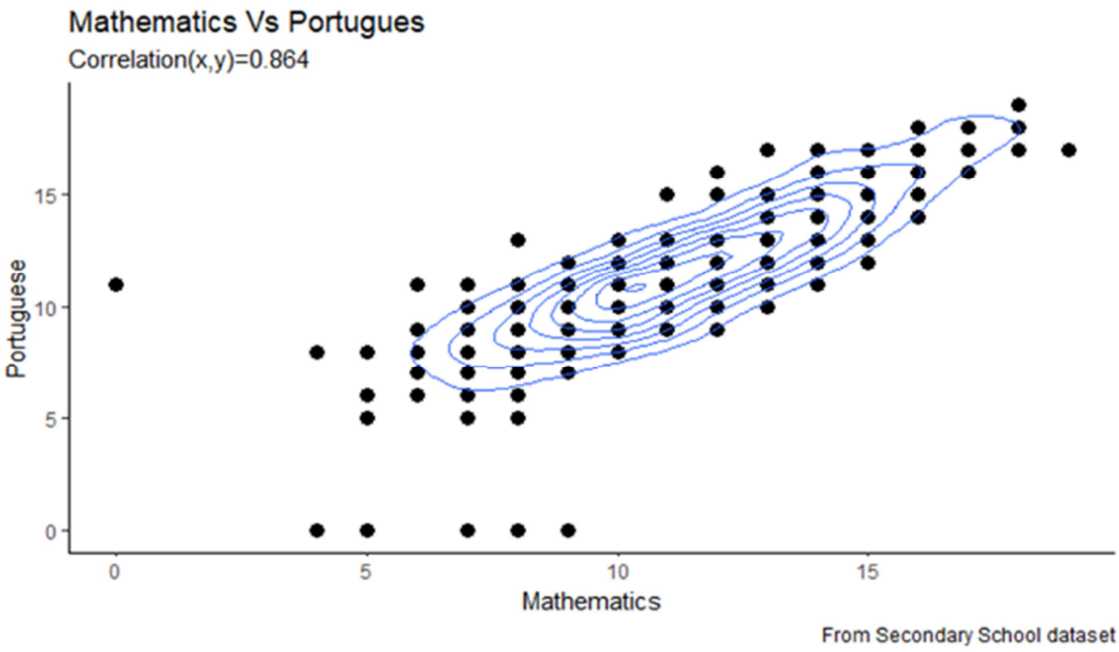

Figure 1. *Scatter plot of Mathematics and Portuguese grades.*

Father job is one of the leading thought of all as evidenced in Table 1 above. We chose as the father of activity corresponded to the two categories of attributes reasonably correlated variables (such as Mathematics and Portuguese value) classified into two agencies (e.g. "teaching" and "other than teaching") and reach the child node of the decision tree. If we plot the two variables of data set one by one as a gain for the class characteristics "daddy work" in two subgroups of "teacher" and "other than teaching", we see that students value of the two groups had a remarkable pattern as evidenced in Figure 2 and Figure 3 below. Figure 1 Scatter plot of Mathematics and Portuguese value have significant differences in their levels in this way, we classify statistics on infant node. Now in every activity father toddler node "teacher" and "other" we repeat the whole method, find the MIC (level dependent) at each level of element and also classifies the daddy "teaching" to the node sub-childe and the father of the "others".

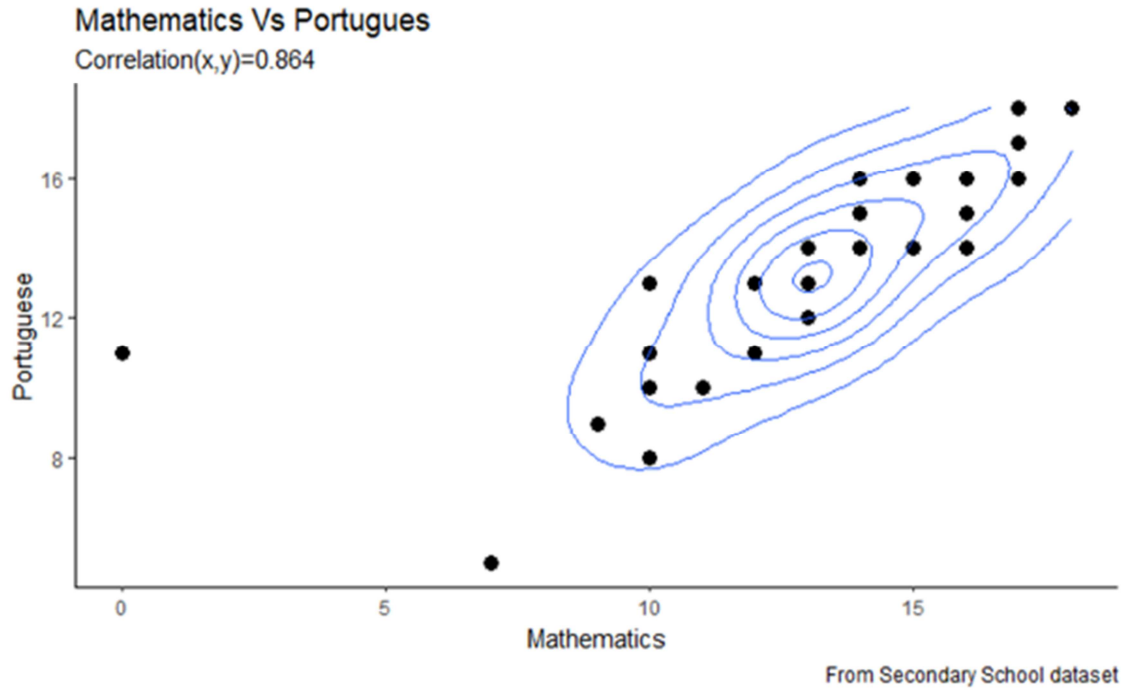

Figure 2. *Fathers job "teacher".*

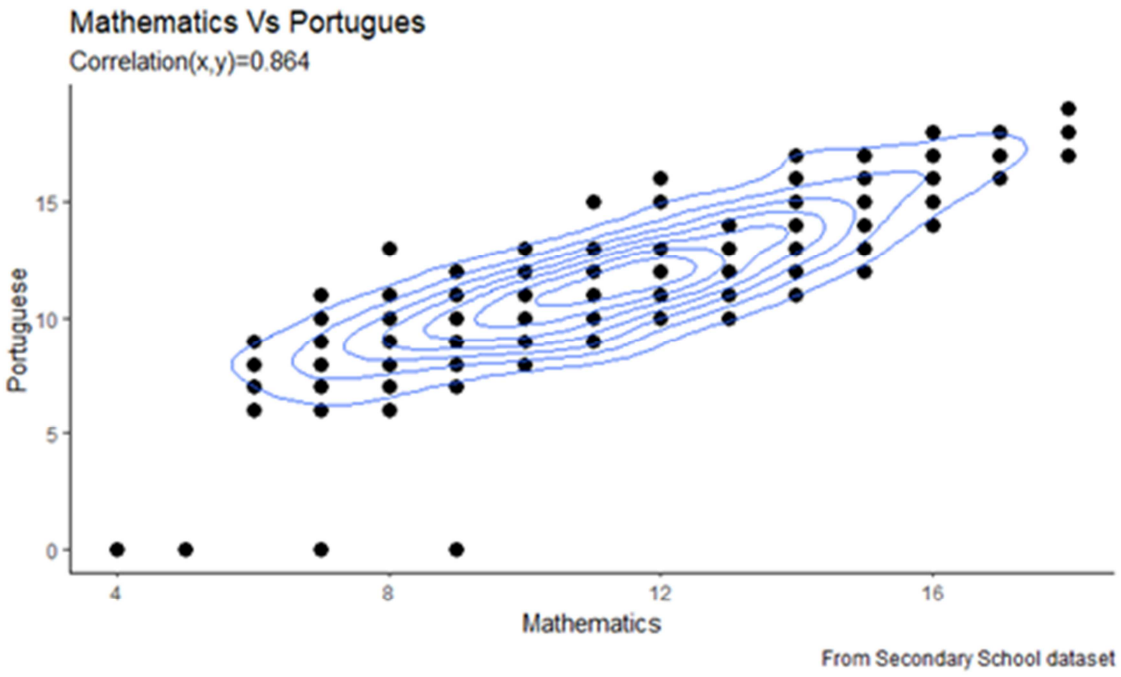

Figure 3. *Fathers job "others".*

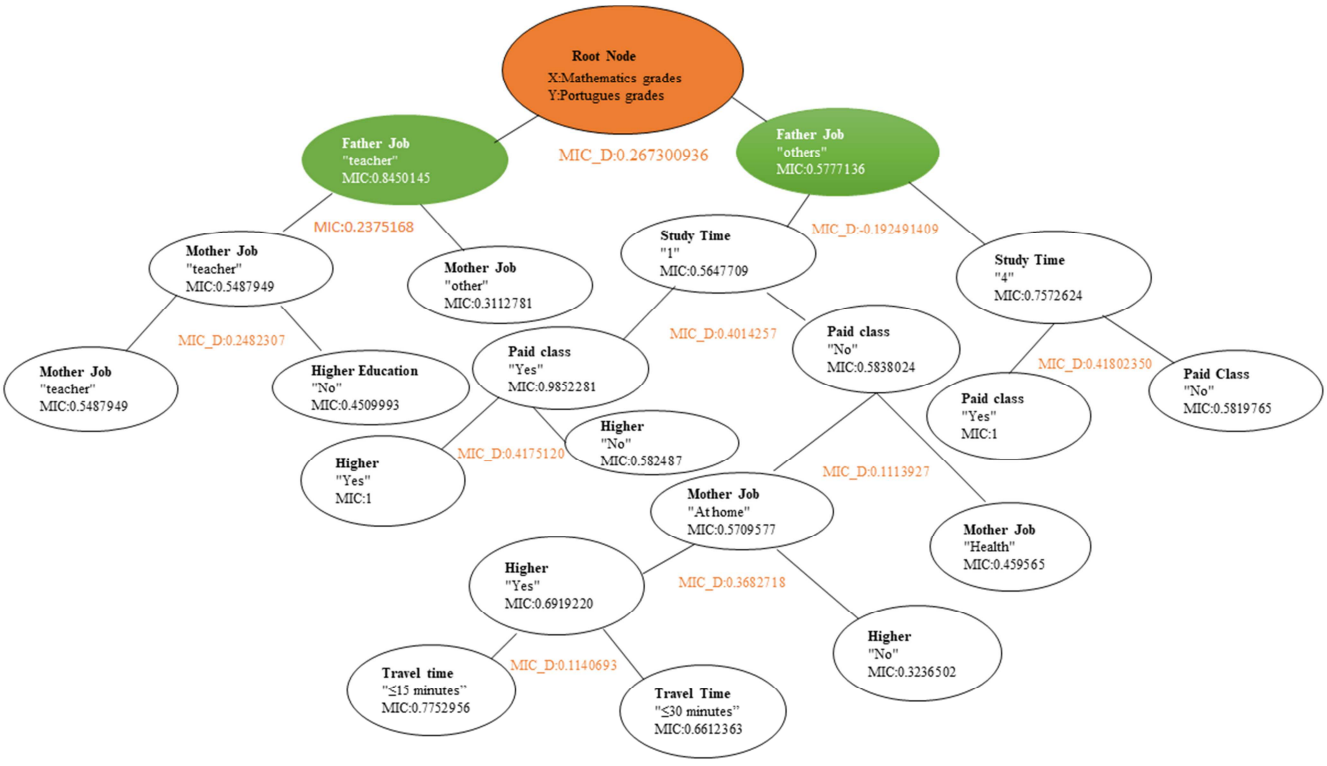

Figure 4. *Two variable decision tree of Students Mathematics and Portuguese grades.*

In the child node "teaching" mother jobs has the highest MIC differences; while in the child node "other" component "study time" has a different maximum MIC. We further classify the child node in the node sub-baby mom job "teacher" and "other" and alternative learning time "2 hours weekly" and "10 hours of weekly". Figures 5, 6, 7 and 8. Appendix A1 is a plot of the students of Mathematics and Portuguese value in accordance with mom job "teacher", the activity of the mother "other" and study time. Similarly, we repeat the whole method on each child and sub-child node and bequeath to develop branches as shown is Figure 8 and stop growing branches of the decision tree in which the difference in MIC at the stage aspect is much less than or identical with a pre-specified value of MIC. After completing the full process on each child and infant sub-nodes we obtain the tree diagram tree shape our choice of two variables.

Under a variable Figure 4 two decision trees for middle school students' overall performance based entirely on Mathematics and Portuguese values together with MIC values in each child node and the difference of MIC between the factor levels. Two variable decision tree is a useful data mining technique which not only efficiently classify the data but also rank the factor according to their importance.

Two variable secondary school student's performance decision tree based on Mathematics and Portuguese grades.

4. Results and Discussion

Section three, provide a brief explanation of how the two variable decision tree for secondary school students from Portugal the use of the degree of dependence (MIC) be constructed thinking about the value of the two central classes (e.g. Mathematics and Portuguese). Here we will discuss the final result directly from the decision tree in the variable information; as set out in Figure 4 above and will get the end of this study are no longer handiest help fathers and mothers, school management to improve the performance of students at the secondary level but they can be useful in making concrete coverage as well. The value of study (e.g. Mathematics and Portuguese) effective- clearly correlated because of this that we would expect the mathematical value and the value of Portuguese as evidenced in Figure 1 section three above. Off the route, it is true if students have a correct knowledge of the native language (e.g. Portuguese) he / she would have a proper understanding of Mathematics; and really perform well in Mathematics. Next among all the factors "father's job" play a very important position in the performance of middle school students as shown in Table 1. Above in section 3. The difference in MIC for factor "father's work" at their level "teaching" and "other" is maximum among all. From Figure 2 above, it is clear that those students whose fathers teaching activities in general they have a better knowledge of both training and their overall performance is above the general level of each class value (e.g. Mathematics and Portuguese). From here we classify a complete data values of the two subjects into two subgroups and gain child-nodes of the tree of our choice where the relationship between the value of the two subgroups is quite one of a kind as shown in Figure 4 above. In infant's node the parent jobs "teaching" there is a strong relationship between a class of students at the secondary stage, to show that the activity of the father "teaching" without delay has an effect on student performance. In child node father job "teaching" the same we learn that the mother's occupation is one of the most influential aspects of both the performance of pupils. Those students whose mother and father each teaching jobs they want to get better schooling and that they have a better performance in the classroom lessons (e.g. Mathematics and Portuguese) as evidenced in Figure 5 Appendix A1. And of course, in the secondary degree there is the important position of the mother and father's education and mother and father work on the overall performance of school students and wishes them have a higher education is no longer just in Portugal but in the whole world. On the other hand, in the Child- node father profession "others" study time is one of the most important issues both of which have a significant impact on student performance. As shown in Figure 4. Section 3. Those students whose father job is other than teaching needs to have at least ten hours of study time in step with a week to do it right on the second class subject (e.g. Mathematics and Portuguese) as evident from Figures 7 and 8 there may be a big difference in the overall performance of students for study time "two hours" and the overall performance of students for study time "ten hours". In both child nodes father jobs "others" after observing the time aspect "paid class" indicates a high impact on overall student performance as shown in Figure 4 Section 3. Which are different from teaching duties father and that a study two hours a week; Mother occupation has an influence on their overall performance. Those students whose mother job is "health" they do not pay for classes at the school while, that the mother work is different from the "health" they want to receive higher education and their performance is also having an impact on the way traveling time from school significantly as the event offered a variable decision tree in Figure 4 on the segment 3. MIC at each child-node and their difference between child-nodes.

5. Conclusion

Schooling playing position lower spine in the betterment of society worldwide. The influence so that one can be a mirror of the long-time period monetary advancement of society. It's much essential element for European society. Data mining (DM) techniques, which allow the extraction of a high level of knowledge of the raw records, and found many hidden opportunities shaping the training domain. In particular, many studies conducted in recent years for the people of Europe using statistical mining enjoyable strategy to improve schooling, policy-making and control of resources also decorate the school.

In this paper, we construct two variable decision tree to correctly classify records and statistics students. We have addressed the most influential factors on the performance of high school students from two main courses (Math and Portuguese) use a variable decision tree records mining methods. Effects displays that, if the data at intervals of the first school and / or both recognized; we can correctly classify information that facilitates predicting overall student performance and get the essential component that plays a major position in the overall performance of high school students for Portugal. As a direct result of this study, the father's occupation is utmost important aspect which has a good size impact on student performance on the secondary stage. For preference to acquire higher education in the EU father teaching jobs and jobs community teaching mothers play an important role. Yang, who is the father of activities apart from teaching their overall performance drastically influenced from observing time as one week. They need a wide range of high hours of study for a week in line with the accurate performance in any instruction (such as Mathematics and Portuguese). For the father of activities

besides teaching, learning time is an important element of the second, paid class came after studying and homework time mothers have a less important function in the performance of pupils. At the secondary stage travel time from school to home also has an influence on overall student achievement but this effect is most. In addition, we need to strengthen research at the stage area of the country to get the beneficial effects to be useful in making choice, especially, in policy making to improve the level of education at the gross root level.

5.1. Implementation and Policy Making

There is a need of high quality to enhance the same type of research at the country level in Portugal apart; all countries in Europe are falling on the tail stops from the table of data because of excessive failure rate of its students. To determine the factors that play a broad function in the performance of students in secondary schools. There is a need for concrete policy-making based on the type of such research at the government level in Europe to overcome the difficulties pupil cost of failure is high.

5.2. Computational Environment and R packages

All calculations reported in this paper are carried out in R (Studio), a programming language that excessive levels of free on-line can be obtained with an effective suite of teeth for statistical and analytical records; is user friendly and bendy. mic anticipated use of "mine" suite ref 14 as defined. R code should have to be given to non-public requests through e mails to the corresponding creator.

Appendix

Appendix A1. Table and Graphs

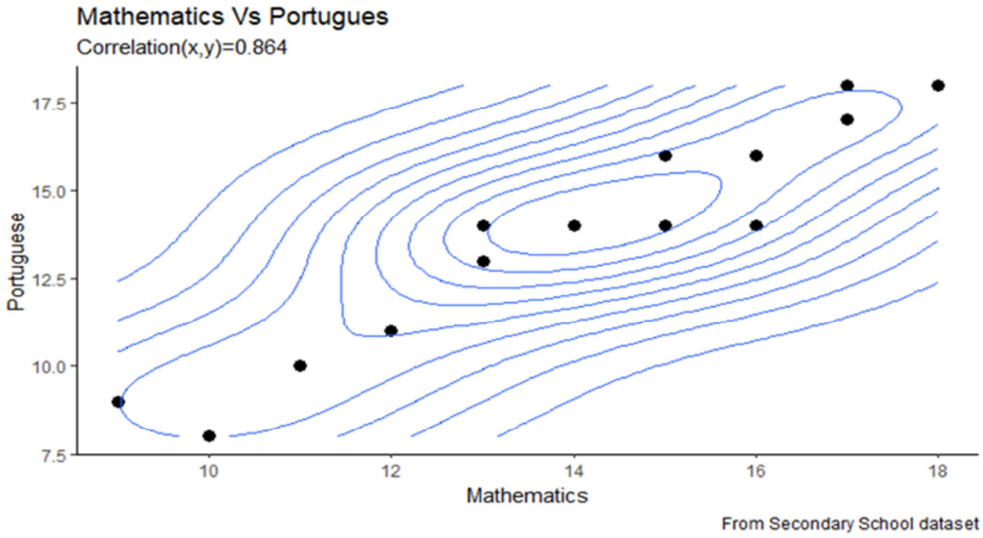

Figure 5. *Father & Mother job "**teacher".***

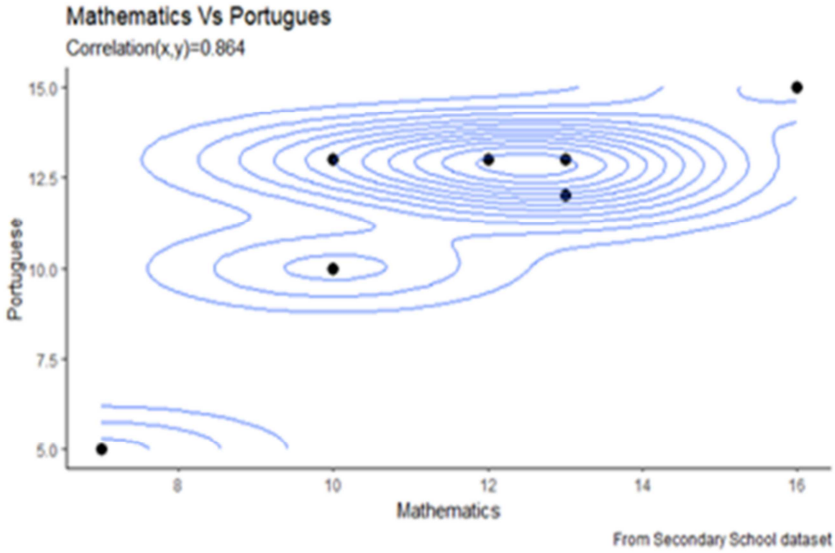

Figure 6. *Father job teacher mother job "other".*

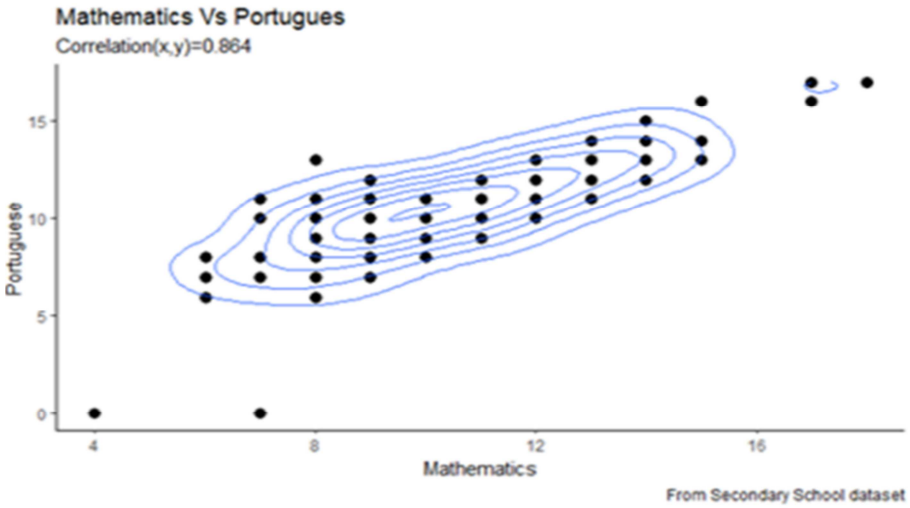

Figure 7. *Study time "≤ 2 hours weekly".*

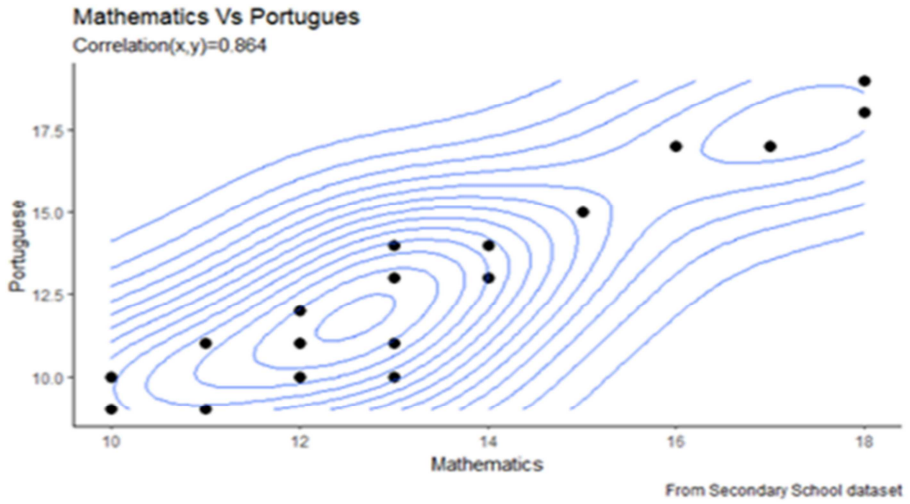

Figure 8. *Study time "> 10 hours weekly".*

Table 2. *Factors having Significant and Insignificant MIC differences at their level.*

S. no	Insignificant Factors	MIC differences at factor Levels	Significant Factors	MIC differences at factors levels
1	Romantic	-0.0038830	Family Relation	-0.20724080
2	Absence	0.03386000	Study Time	-0.26240197
3	School sup	0.04133134	Paid class	-0.12804757
4	Internet	0.00690000	Higher Education	0.14823272
5	Go out	0.07106400	Father Job	0.267300936
6	Weekly alc	-0.04729313	Mother Job	-0.25862190
7	Family sup	-0.0898155	Failures	-0.26240190
8	Daily alc	0.01674400	Travel Time	0.203002284
9	Nursery	0.10467000		
10	Reason	0.05050000		
11	Guardian	0.01150000		
12	Sex	0.01900000		
13	Father Edu	-0.0443201		
14	Family size	0.07687870		
15	Health	-0.04729320		
16	Pstatus	-0.01265022		
17	Address	0.011460212		
18	Mother Edu	-0.017313610		
19	Free Time	-0.083511146		

Appendix A2. Data Sources and Brief Explanation

Table 3. *Secondary school students related variables and their description.*

Variables	Description
Sex	student's (binary: female or male)
Age	student's age (numeric: from15to22)
School	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
Address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together apart)
Medu	education (numeric: from0to4a)
Mjob Fedu	mother's job (nominal[b]) father's education (numeric: from0 to 4[a])
Fjob	father's job (nominalb)
Gurdain	student's guardian (nominal: mother, father or other)
Famsize	family size (paired: ≤ 3 or >3)
Famrel	nature of family connections (numeric: from 1 – extremely terrible to 5 – astounding)
Reason	motivation to pick this school (ostensible: near and dear, school notoriety, course inclination or other)
Travel time	home to school travel time (numeric: 1 – <15min., 2–15 to30 min., 3 – 30 min to1 hour or 4 – >1 hour).
Study time	Weekly study time (numeric: 1–<2hours, 2–2to5hours, 3–5to10hoursor4–>10hours)
Failures	number of past class disappointments (numeric: n if 1 ≤ n <3, else4)
Schoolsup	extra instructive school support (double: indeed, or no)
Famsup	family instructive help (parallel: indeed, or no)
Activities	activities extra-curricular exercises (parallel: indeed, or no)
Paid class	extra paid classes (parallel: indeed, or no)
Internet	Web access at home (parallel: indeed, or no)
Nursery	gone to nursery school (parallel: indeed, or no)
Higher	wants to take higher education (binary: yes, or no)
Romantic	with a romantic relationship (binary: yes, or no)
Free time	leisure time after school (numeric: from 1 – exceptionally low to 5 – very high)
Gout	going out with companions (numeric: from 1 – extremely low to 5 – very high)
Walc	going out with companions (numeric: from 1 – extremely low to 5 – very high)
Dalc Health	current wellbeing status (numeric: from 1 – extremely awful to generally excellent)
Absences	number of school absences (numeric: from 0 to93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to20)

References

[1] Eurostat., (2007). Early school-leavers. http://epp.eurostat.ec.europa.eu/.

[2] Mejer. L., Turchetti.P., and Gere. E., (2011). Trends in European Education during the Last Decade. http://epp.eurostat.ec.europa.eu/.

[3] Turban. E., Sharda. R., Aronson. J., and King. D., (2007). Business Intelligence, A Managerial Approach. Prentice-Hall.

[4] Ma. Y., Liu. B., Wong C., Yu. P., and Lee S., (2000). Targeting the Right Students Using Data Mining. In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464.

[5] Luan. J., (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 113, 17–36.

[6] Minaei-Bidgoli. B., Kashy. D., Kortemeyer G., and Punch. W., (2003). Predicting Student Performance: an application of data mining methods with an educational web-based system. In Proc. of IEEE Frontiers in Education. Colorado, USA, 13–18.

[7] Kotsiantis. S., Pierrakeas C., and Pintelas. P., (2004). Pre-dicting Students' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence (AAI), 18, no. 5, 411–426.

[8] Pardos. Z., Heffernan N., Anderson. B., and Heffernan. C., (2006). Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. In Proc. of 8th Int. Conf. On Intelligent Tutoring Systems. Taiwan.

[9] Cortez. P, and A. Silva. A., (2008) Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, EUROSIS, ISBN 978-9077381-39-7.

[10] Patel. B. N., Prajapati. S. G., and Lakh aria. K. I., (2012). Efficient Classification of Data Using Decision Tree. Bunfring International Journal of Data Mining, vol. 2, no. 1, pp. 6-12.

[11] Wang. L. M., Li. X. L., Cao. C. H., and Yuan. S. M., (2006). Combining Decision Tree and Naïve Bayes for Classification. Knowledge-Based Systems, vol. 19, no. 7, pp. 511–515.

[12] Aitkenhead. M. J., (2008). A Co-Evolving Decision Tree Classification Method, Expert Systems with Applications, vol. 34, no. 1, pp. 18–25.

[13] Kraskov. A., Stogbauer. H., and Grass Berger. P., (2004). Estimating Mutual Information. Phys Rev E Stat Nonlin Soft Matter Phys 69 (6 Pt 2): 066138.

[14] Kinney. J. B., and Gurinder. S. A (2014). Equitability, mutual information, and the Maximal Information Coefficient. PNAS, vol. 111, no. 9, pp. 3354–3359.

[15] Reshef. D. N., et al. (2011). Detecting Novel Associations in Large Data Sets. Science 334 (6062): 1518-1524.

[16] Reshef. D. N., Reshef. Y., Mitzenmacher. M., and Sabeti. P., (2013) Equitability Analysis of the Maximal Information Coefficients with Comparisons. arXiv: 1301.6314v1 [cs. LG].

[17] Hastie. T., Tibshirani. R., and Friedman. J. H (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction Springer Verlag, New York.

[18] Filose M, et al. (2013). Minerva: Maximal Information-Based Nonparametric Exploration R package for Variable Analysis version 1.3 URL http://www.r-project.org, http://mpba.fbk.eu/cmine.