Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.13, No.5, 2022

Determinants and Prediction of Secondary Students Performance in Mathematics in Portugal Using Machine Learning

Md Nesar Uddin Sorkar [1*] Md. Roquib Uddin Sorkar [2]
1. Shahjalal University of Science and Technology, Sylhet, Bangladesh
2. Southern University Bangladesh, Chittagong, Bangladesh
* E-mail of the corresponding author: nesarsust@gmail.com

Abstract
Mathematics brings order and prevents chaos in our lives. Power of reasoning, inventiveness, abstract or spatial thinking, critical thinking, problem-solving abilities, and even excellent communication skills are some of the attributes that mathematics fosters. The purpose of this research is to identify the factors that impact students' mathematics performance. Simultaneously, this study tries to predict student's mathematics performance. The secondary information was gathered from https://archive.ics.uci.edu/ml/datasets/student+performance#. This study uses a variety of methodologies such as percentage distribution, association tests, association rule mining, Decision tree, Random Forest, Support Vector Machine, Naive Bayes, K-Nearest-Neighbors, Linear Discriminant Analysis, Neural Network, and Logistic Regression. Students whose father's education is higher, mother's education is higher, mothers' occupation services, who rarely go out with friends, and who did not fail the previous class have considerably better mathematics performance. Students whose father has a primary education, who spend a lot of time with friends, drink alcohol during the workday, and live in rural areas have poor mathematics performance. Furthermore, in this case, the best classifier for prediction is a Neural Network. Therefore, the performance of students in mathematics depends on the characteristics of the students as well as the characteristics of their parents and place of residence. The government and other NGOs must step forward to educate the country's people about these critical factors.
Keywords: Determinant, Prediction, Student, Performance, Machine, Learning.
DOI: 10.7176/CEIS/13-5-03
Publication date: October 31st 2022

1 Background

1.1 Introduction of the Study: Many students perceive mathematics as a difficult subject to learn and master in class (Siaw *et al.*, 2020). The majority of students struggle and eventually fail the subject (Mazana, Montero and Casmir, 2020). Strong mathematics performance is largely dependent on the gender of students, with male students performing much better than female students (Asampana, Kassim Nantomah and Ayagikwaga Tungosiamu, 2017). Students' performance in mathematics was significantly influenced by emotional intelligence, self-esteem, and self-efficacy (Ugwuanyi, Okeke and Asomugha, 2020). Mathematics performance is highly influenced by teacher classroom activities (Arends, Winnaar and Mosimege, 2017). Missing too many lectures and living in a crowded household are the two criteria that have the greatest detrimental impact on student performance (Harb and El-Shaarawi, 2006). Students' mathematics performance is directly influenced by family support, motivation, involvement, and collaboration between students and teachers (Walde, 2019).

Students' performance in class and on the forum is positively associated with the features of the advising network. Almost all students' performance is negatively connected with adversarial characteristics (Yang and Tang, 2003). Academic ability, testability, time management, and test worry were all found to be significant in students' academic performance (Talib and Sansgiry, 2012). Students who had toothaches were nearly four times more likely to have a low GPA (Seirawan, Faust and Mulligan, 2012). From high school on, the mathematics subject area interacts significantly with both student age and English skills — negatively for age and positively for English skills (Ketkaew and Naruetharadhol, 2015).

1.2 Research Objective: The goal of this study is to figure out what factors influence students' mathematics performance. Simultaneously, this research is attempting to predict student mathematics performance.

1.3 Statement of Problem: The majority of students are terrified of mathematics. As a result, students frequently underperform in mathematics. But, mathematics is a crucial subject for students. It will be quite difficult for them to achieve success if they do not have a thorough understanding of mathematics. As a result, determining the factors that contribute to students' low mathematics performance has become essential.

1.4 Scope of Research: Here this data has been collected from the students of Portugal so this study will cover the whole of Portugal.

1.5 The rationale of the Study: No one has used data mining to find out the determinants of students' performance in mathematics, so here is an attempt to do it.

1.6 Significance of the Study: As a result of this study, students in Portugal will be able to know which factors are causing their poor performance in mathematics. Therefore, students of Portugal will be more beneficiaries

through this study.

2 Methods

2.1 Data source: The secondary information was gathered from https://archive.ics.uci.edu/ml/datasets/student+performance#.

2.2 Study Area: These data were taken from secondary students of Gabriel Pereira and Mousinho da Silveira schools in Portugal.

2.3 Outcome variable: In this study, the result in mathematics is the outcome variable. This variable has been created by dividing the final grade variable into two categories. If the value of the final grade is more than 10, it is considered a high result and if the value of the final grade is less than 10, it is considered a low result.

2.4 Predictor variables: Here Sex, Romantic relationship, Going out with friends, Workday alcohol consumption, Weekend alcohol consumption, Current health status, School, Free time after school, Past class failures, Number of school absences, attended nursery school, School travel time, Reason to choose this school, wants to take higher education, extracurricular activities, Weekly study time, extra paid classes, Extra educational support, Residence, Family size, parents cohabitation status, Mothers education, Fathers education, Mothers job, Father's job, Students guardian, Family educational Support, Quality of family relationships, Internet access at home variables have been used as independent variables.

2.5 Statistical Analysis: Here, different kinds of methodologies such as percentage distribution, association tests, association rule mining, Decision tree, Random Forest, Support Vector Machine, Naive Bayes, K-Nearest-Neighbors, Linear Discriminant Analysis, Neural Network, and Logistic Regression were used. Optimal accuracy, optimal sensitivity, optimal specificity, the optimal area under the curve, and optimal Brier Score have been extracted using out-of-sample cross-validation for each method. The data was analyzed using SPSS 23, R 3.6.3, and Microsoft Excel.

3 Results

3.1 Percentage Distribution

3.1.1 Percentage Distribution of Students Characteristics:

Table-1 illustrates the various characteristics of students. It can be seen that around 53% of the students are female and 47% are male. Around 34% of students are in a romantic relationship, while about 12% of students have poor health.

Table-1: Percentage Distribution of Students Characteristics

Characteristics		Frequency(%)
Sex	female	208(52.66%)
	male	187(47.34%)
Romantic relationship	no	263(66.58%)
	yes	132(33.42%)
Going out with friends	Very low	23(5.82%)
	low	103(26.08%)
	Middle	130(32.91%)
	High	86(21.77%)
	Very High	53(13.42%)
Workday alcohol consumption	Very low	276(69.87%)
	low	75(18.99%)
	Middle	26(6.58%)
	High	9(2.28%)
	Very High	9(2.28%)
Weekend alcohol consumption	Very low	151(38.23%)
	low	85(21.52%)
	Middle	80(20.25%)
	High	51(12.91%)
	Very High	28(7.09%)
Current health status	Very bad	47(11.9%)
	bad	45(11.39%)
	Middle	91(23.04%)
	Good	66(16.71%)
	Very Good	146(36.96%)

3.1.2 Percentage Distribution of Students Academic Characteristics:

Table-2 highlights the academic characteristics of the students. From **Table-2** we can say that about 5% of the

students have very little free time after school. About 4% of students have failed three times. Here we also see that about 46% of the students have bad results.

Table-2: Percentage Distribution of Students Academic Characteristics

Characteristics		Frequency(%)
School	Gabriel Pereira	349(88.35%)
	Mousinho da Silveira	46(11.65%)
Free time after school	Very low	19(4.81%)
	low	64(16.2%)
	Middle	157(39.75%)
	High	115(29.11%)
	Very High	40(10.13%)
Past class failures	No failure	312(78.99%)
	One time	50(12.66%)
	Two times	17(4.3%)
	Three times	16(4.05%)
Number of school absences	0-5	249(63.04%)
	6-10	80(20.25%)
	>=11	66(16.71%)
attended nursery school	no	81(20.51%)
	yes	314(79.49%)
School travel time	<15 min	257(65.06%)
	15 to 30 min	107(27.09%)
	30 min. to 1 hour	23(5.82%)
	>1 hour	8(2.03%)
Reason to choose this school	course preference	145(36.71%)
	close to home	109(27.59%)
	other	36(9.11%)
	school reputation	105(26.58%)
wants to take higher education	no	20(5.06%)
	yes	375(94.94%)
Extracurricular activities	no	194(49.11%)
	yes	201(50.89%)
Weekly study time	<2 hours	105(26.58%)
	2 to 5 hours	198(50.13%)
	5 to 10 hours	65(16.46%)
	>10 hours	27(6.84%)
extra paid classes	no	214(54.18%)
	yes	181(45.82%)
Extra educational support	no	344(87.09%)
	yes	51(12.91%)
Result	Low	186(47.09%)
	High	209(52.91%)

3.1.3 Percentage Distribution of Students Family Characteristics:

Table-3 shows that about 22% of the students live in rural areas. About 20% of students' parents have primary education. The guardian of most students is their mothers, and about 16% of students do not have an internet connection at home.

Table-3: Percentage Distribution of Students Family Characteristics

Characteristics		Frequency (%)
Residence	rural	88(22.28%)
	urban	307(77.72%)
Family size	>3	281(71.14%)
	<= 3	114(28.86%)
parents cohabitation status	apart	41(10.38%)
	Living together	354(89.62%)
Mothers education	none	3(0.76%)
	primary education	59(14.94%)
	5th to 9th grade	103(26.08%)
	secondary education	99(25.06%)
	higher education	131(33.16%)
Fathers education	none	2(0.51%)
	primary education	82(20.76%)
	5th to 9th grade	115(29.11%)
	secondary education	100(25.32%)
	higher education	96(24.3%)
Mothers job	At home	59(14.94%)
	care-related	34(8.61%)
	other	141(35.7%)
	services	103(26.08%)
	teacher	58(14.68%)
Fathers job	At home	20(5.06%)
	care-related	18(4.56%)
	other	217(54.94%)
	services	111(28.1%)
	teacher	29(7.34%)
Students guardian	father	90(22.78%)
	mother	273(69.11%)
	other	32(8.1%)
Family educational support	no	153(38.73%)
	yes	242(61.27%)
Quality of family relationships	Very bad	8(2.03%)
	bad	18(4.56%)
	Middle	68(17.22%)
	Good	195(49.37%)
	Very Good	106(26.84%)
Internet access at home	no	66(16.71%)
	yes	329(83.29%)

3.2 Association Test

3.2.1 Association between Result and Students Characteristics:

Table-4 shows that the variable going out with friends is significant because its p-value is less than 0.05. The p-value of workday alcohol consumption and weekend alcohol consumption variables are less than 0.1, so they are approximately significant. As a result, going out with friends, workday alcohol consumption, and weekend alcohol consumption variables are associated with the student's mathematics result.

Table-4: Association between Result and Students Characteristics

Characteristics	χ^2	p-value
Sex	1.752	0.186
Romantic relationship	0.005	0.942
Going out with friends	12.839	0.012
Workday alcohol consumption	8.029	0.091
Weekend alcohol consumption	8.587	0.072
Current health status	4.856	0.302

3.2.2 Association between Result and Students Academic Characteristics:

Similarly, **Table-5** shows that free time after school, past class failures, school travel time, wants to take higher education, and extra educational support variables are associated with students' mathematics results.

Table-5: Association between Result and Students Academic Characteristics

Characteristics	χ^2	p-value
School	2.313	0.128
Free time after school	11.004	0.027
Past class failures	42.628	0.00
Number of school absences	1.128	0.569
attended nursery school	0.168	0.682
School travel time	7.754	0.051
Reason to choose this school	1.954	0.582
wants to take higher education	7.82	0.005
Extracurricular activities	0.001	0.976
Weekly study time	5.045	0.169
extra paid classes	0.57	0.450
Extra educational support	6.506	0.011

3.2.3 Association between Result and Students Family Characteristics:

In the same way, **Table-6** also presents that residence, mothers' education, fathers' education, mothers' job, internet access at home variables are associated with the students' mathematics results.

Table-6: Association between Result and Students Family Characteristics

Characteristics	χ^2	p-value
Residence	3.814	0.051
Family size	0.501	0.479
parents cohabitation status	0.86	0.354
Mothers education	14.489	0.006
Fathers education	16.729	0.002
Mothers job	15.867	0.003
Fathers job	3.534	0.473
Students guardian	2.119	0.347
Family educational support	0.277	0.598
Quality of family relationships	3.415	0.491
Internet access at home	3.011	0.083

3.3 Variable Extraction for Association Rule Mining and Prediction

From **Table-4** to **Table-6** it can be seen that going out with friends, workday alcohol consumption, weekend alcohol consumption, free time after school, past class failures, school travel time, wants to take higher education, extra educational support, residence, mothers education, fathers education, mothers job and internet access at home variables are associated with student's mathematics result.

So it can be said that the mathematics result of the students may be related to these variables. For this reason, only these significant variables have been used in association rule mining and prediction algorithms. As a result, the prediction will be better and the result of association rule mining will also be better.

3.4 Influence of Different Characteristics

3.4.1 Characteristics Influencing Students High Results in Mathematics

According to association rule mining, **Figure-1** shows that the color of the rule-1 bubble is very dark. So it can be said that among the students whose free time is low after school, their result in mathematics is much better. Similarly, among the students whose fathers' education is higher, mothers' education is higher, mothers' occupation services, their results in mathematics is also better. At the same time, those who rarely go out with friends did not fail in the previous class, their result in mathematics is also much better.

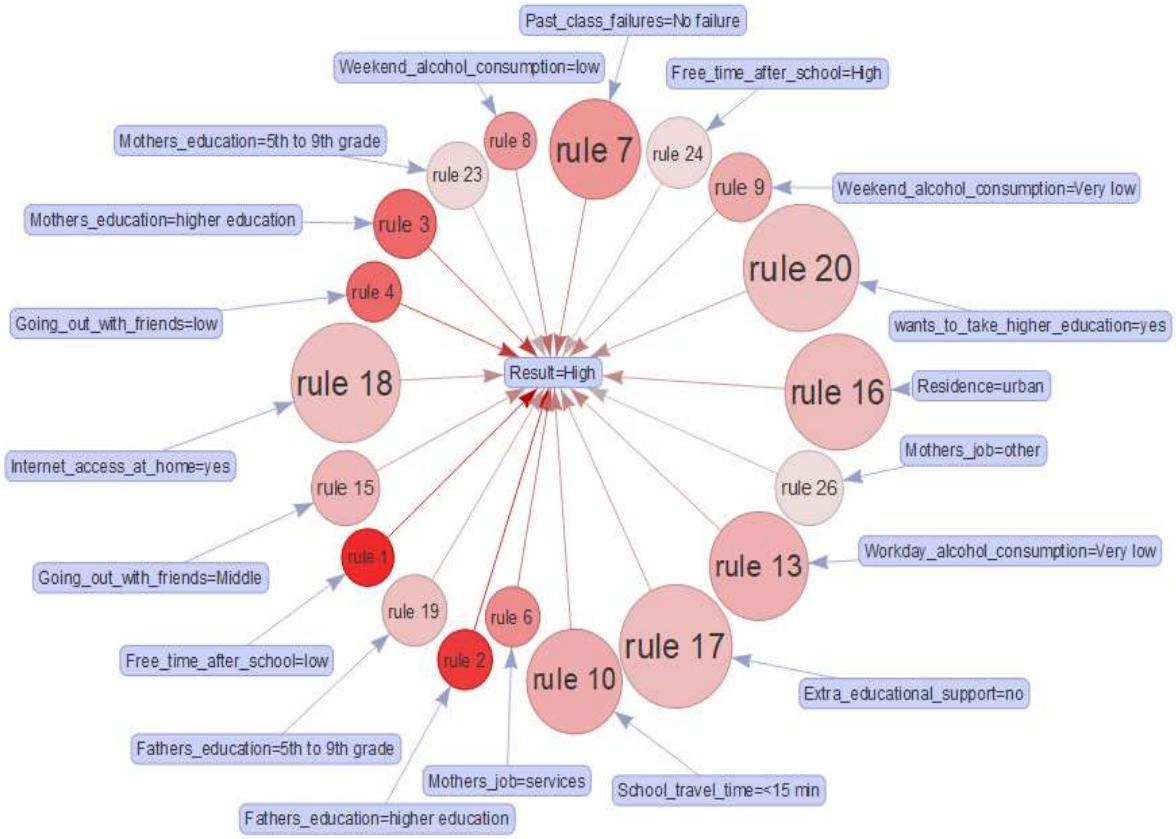


Figure-1: Characteristics Influencing Students High Results in Mathematics

3.4.2 Characteristics Influencing Students Low Results in Mathematics

According to association rule mining, **Figure-2** shows that among the students whose father's education is primary, who go around a lot with friends have poor results in mathematics. Besides, those who consume alcohol on a workday and live in rural areas also have poor results in mathematics.

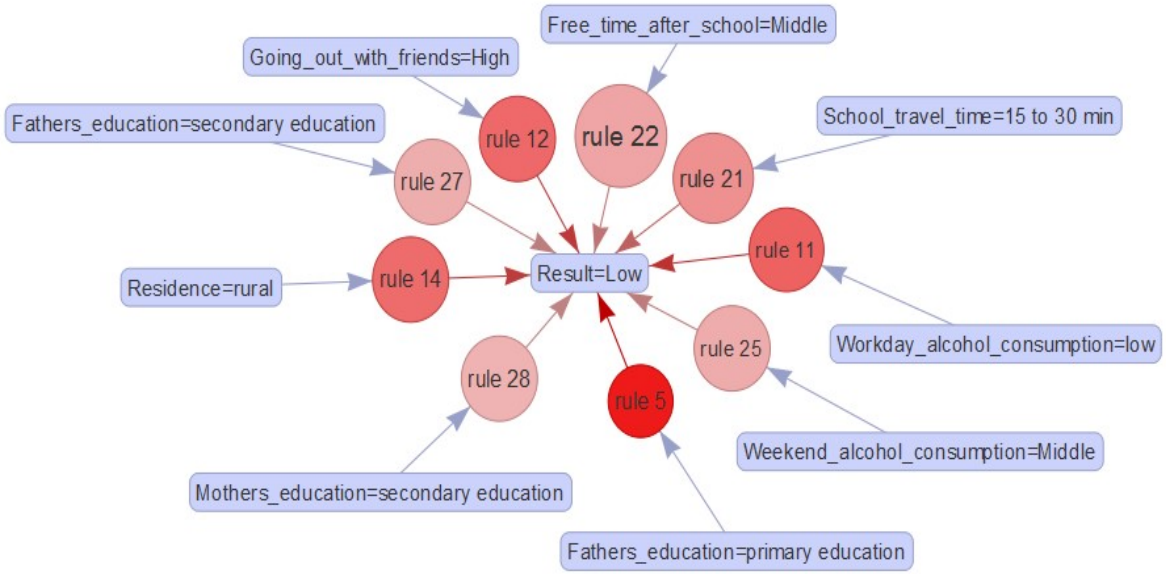


Figure-2: Characteristics Influencing Students Low Results in Mathematics

3.5 Prediction Algorithms Characteristics

3.5.1 Optimal Accuracy: The accuracies of the Decision Tree, Random Forest, Support Vector Machine, Naive Bays, K-Nearest-Neighbors, Linear Discriminant Analysis, Neural Network, and Logistic Regression classifiers are approximately 59.18 %, 63.39 %, 62.33 %, 65.65 %, 57.81 %, 64.74 %, 73.94 %, and 62.56 %, respectively, as shown in **Figure-3(a)**. Here, the accuracy of the Neural Network is the highest, hence it is the

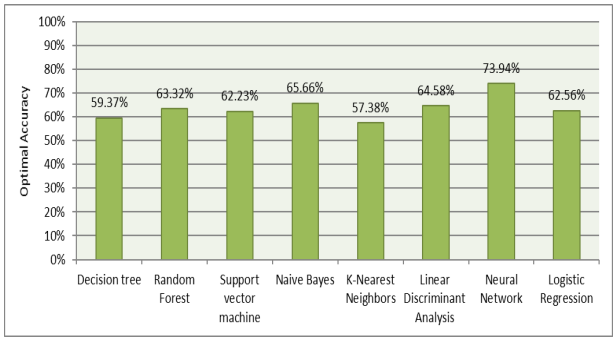
best classifier in terms of accuracy value.

3.5.2 Optimal Sensitivity: Sensitivity refers to a test's ability to detect true events as true. The Neural Network has the highest sensitivity, as seen in **Figure-3(b)**. As a result, the neural network is the best classifier in terms of sensitivity value.

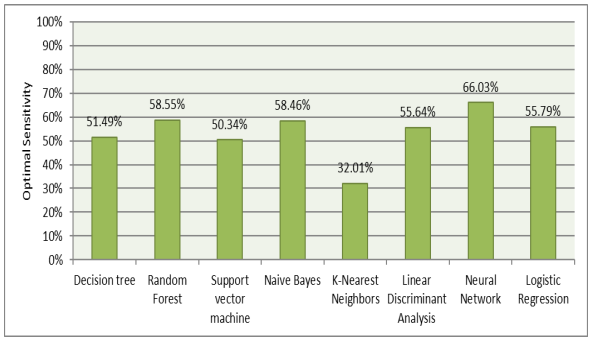
3.5.3 Optimal Specificity: Specificity refers to a test's ability to detect false events as false. K-Nearest-Neighbors specificity is the highest, as seen in **Figure-3(c)** (about 80.64%). So, in terms of the value of specificity, K-Nearest-Neighbors is the best classifier.

3.5.4 Optimal Area Under Curve (AUC): Figure-3(d) demonstrates that the Neural Network has the largest area under the curve. The value of AUC is approximately 73.98%. As a result, the Neural Network is the best classifier here as well.

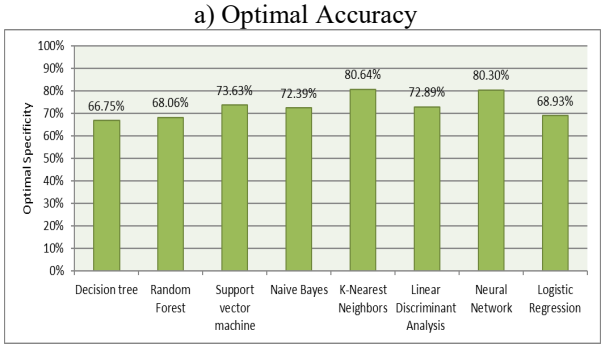
3.5.5 Optimal Brier Score: A brier score is a way to check whether a probability prediction is accurate. A probability prediction is a prediction about an upcoming event. The technique is better if the Brier Score is lower. The neural network has the lowest Brier Score (0.228) in **Figure-3(e)**. So, the Neural Network is the best classifier in this case.



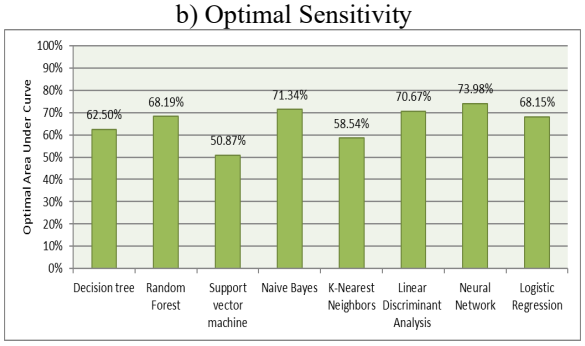
a) Optimal Accuracy



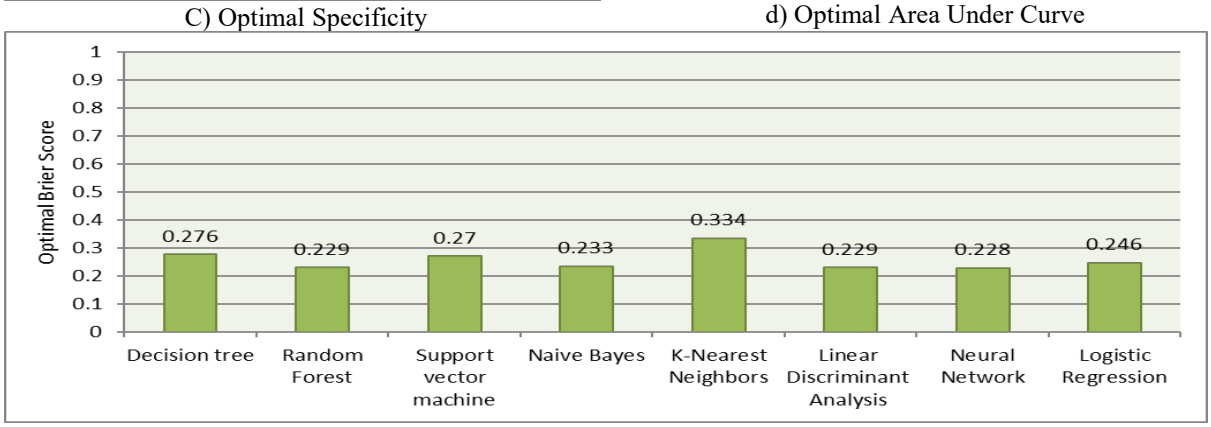
b) Optimal Sensitivity



C) Optimal Specificity



d) Optimal Area Under Curve



e) Optimal Brier Score

Figure-3: Prediction Algorithms Characteristics

3.5.6 Comparison of Different Algorithms

In most cases, the value of accuracy, sensitivity, specificity, and area under the curve of Neural Network is higher than the remaining seven methods, as shown in **Table-7**. Neural Network has the lowest Brier Score of the other seven approaches. Hence, the Neural Network is the best classifier in this case.

Table-7: Comparison of Different Algorithms

Algorithms	Optimal Accuracy	Optimal sensitivity	Optimal specificity	Optimal AUC	Optimal Brier Score
Decision tree	59.37%	51.49%	66.75%	62.50%	0.276
Random Forest	63.32%	58.55%	68.06%	68.19%	0.229
Support Vector Machine	62.23%	50.34%	73.63%	50.87%	0.27
Naiv Bays	65.66%	58.46%	72.39%	71.34%	0.233
K-Nearest Neighbors	57.38%	32.01%	80.64%	58.54%	0.334
Linear Discriminant Analysis	64.58%	55.64%	72.89%	70.67%	0.229
Neural Network	73.94%	66.03%	80.30%	73.98%	0.228
Logistic Regression	62.56%	55.79%	68.93%	68.15%	0.246

4 Discussion

The walking distance to school, the child's sex, the parent's or guardian's educational status, dietary levels, late entry and repetition at school, and the language is spoken at home have all been identified as determinants of academic performance (Garikai, 2010). This research is similar to this one. In this study, students' math performance is influenced by their parents' educational levels and the mother's profession, etc.

5 Strengths and Limitation

Here, data has been collected from students of only two schools in Portugal. If it is possible to collect data from more schools, the result of this study may be better.

6 Conclusion

Students whose father's education is higher, mother's education is higher, mothers' occupation services, who rarely go out with friends, and did not fail in the previous class their mathematics results are much better. On the other hand, for students whose father's education is primary, who spend a lot of time with friends, drink alcohol during the workday, and live in rural areas their mathematics result is poor. Besides, In this case, Neural Network is the best classifier for prediction.

7 Recommendation

The government and various NGOs need to come forward to make the people of the country aware of these significant variables. As a result, the fear of mathematics will be reduced among the students.

Acknowledgments

We would like to thank the UCI Machine Learning Repository for using their data.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Arends, F., Winnaar, L. and Mosimege, M. (2017) 'Teacher classroom practices and mathematics performance in South African schools: A reflection on TIMSS 2011', *South African Journal of Education*. doi: 10.15700/saje.v37n3a1362.

Asampana, G., Kassim Nantomah, K. and Ayagikwaga Tungosiamu, E. (2017) 'Multinomial Logistic Regression Analysis of the Determinants of Students' Academic Performance in Mathematics at Basic Education Certificate Examination', *Higher Education Research*.

Garikai, B. W. (2010) 'Determinants Of Poor Academic Performance', *Articlesbase*.

Harb, N. and El-Shaarawi, A. (2006) 'Factors Affecting Students' Performance', *Journal of Business Education*.

Ketkaew, C. and Naruetharadhol, P. (2015) 'Determinants of International College Student's Performance in Mathematics', *Procedia - Social and Behavioral Sciences*. doi: 10.1016/j.sbspro.2015.06.460.

Mazana, M. Y., Montero, C. S. and Casmir, R. O. (2020) 'Assessing Students' Performance in Mathematics in Tanzania: The Teacher's Perspective', *International Electronic Journal of Mathematics Education*. doi: 10.29333/iejme/7994.

Seirawan, H., Faust, S. and Mulligan, R. (2012) 'The impact of oral health on the academic performance of disadvantaged children', *American Journal of Public Health*. doi: 10.2105/AJPH.2011.300478.

Siaw, E. S. *et al.* (2020) 'Understanding the Relationship Between Students' Mathematics Anxiety Levels and Mathematics Performances at the Foundation Level', *Journal of Education and Learning*. doi:

10.5539/jel.v10n1p47.

Talib, N. and Sansgiry, S. S. (2012) 'Determinants of Academic Performance of University Students', *Pakistan Journal of Psychological Research*.

Ugwuanyi, C. S., Okeke, C. I. O. and Asomugha, C. G. (2020) 'Prediction of learners' mathematics performance by their emotional intelligence, self-esteem and self-efficacy', *Cypriot Journal of Educational Sciences*. doi: 10.18844/cjes.v15i3.4916.

Walde, G. S. (2019) 'Hierarchical linear model to examine determinants of students' mathematics performance', in *Journal of Physics: Conference Series*. doi: 10.1088/1742-6596/1176/4/042088.

Yang, H. L. and Tang, J. H. (2003) 'Effects of social network on students' performance: A web-based forum study in Taiwan', *Journal of Asynchronous Learning Network*. doi: 10.24059/olj.v7i3.1848.