Influence of Feature Selection on the Prediction of Student Performance

Raya N. Ismail
Department of Computer Sciences, College of Computer & math. Science, Tikrit University, Iraq
Email: raya_computer@tu.edu.iq

Armanesa Naaman Hasoon
Department of Computer Sciences, College of Computer & math. Science, Tikrit University, Iraq

Israa Rafaa Abdulqader

ABSTRACT

In the process of students evaluation most academic institutions assume that the performance of student is the major criteria. Machine learning offers different techniques used in several fields of education including student performance. This paper presents analytic study to find the most affective attributes related to student academic performance by applying classification algorithms on a collected student`s data. The data collected from Computer science department in Tikrit University-Iraq. The attributes labeled into four categories (personal, family, study, and online activities) then a combination of classification models tested on each type of the attributes. This study aims to give the academic educators good understanding of the obstacles facing their student and could affect their grades. The subset of "study-attributes" resulted best accuracy in all models.

Keywords: Student performance, Machine learning, Feature selection

1. Introduction

Huge number of student's information become the reason why researches in the field of education increased so fast. In order to support educational institutions in recognizing their students' performance both classification and clustering techniques used to evaluate student's performance in academia. The use of academic student data set from Kerala, India showed that this combination achieves superior results in prediction accuracy of student performance [1]. Regression with classification algorithm can produce good result in the prediction process. This combination is used with a student data set from UCI repository the result showed that Random Forest classifier obtain advanced prediction accuracy than regression [2]. In the process of comparison, four models used to select attributes with different classification algorithms to predict student performance the result showed that minimum redundancy and maximum relevance gave the most effective set of attributes, which can be significant in the classification [3]. R.Bertolin et al. integrated feature selection methods with cross validation a significant value for Area Under the Curve achieved when using Fisher `s Scoring Algorithm with correlation Attribute Evaluation [4]. M. Zaffar used two different student datasets to analyze the performance of classification models with Filter feature selection algorithms. To help educators finding the most relevant features to predict student performance. The study presented a difference percentage between the accuracy before and after FS usage [5]. To determine the most powerful factors that can affect the performance of classifiers, which enhance the prediction performance a new attribute selection method, is produced. The results showed that the use of attributes selected by CHIMI improved classifiers accuracy [6]. Also to develop a prediction system for student performance a InfoGainAttributeEval used as a filter attribute selection method and WrapperSubsetEval as a wrapper attribute selection algorithm to select relevant features and the resulted set of features implemted to classify a data set with Random Forest classifier [7]. A student`s data collected from University of Technology Thanyaburi used to evaluate classifiers performance. Neural Network model with mutual information algorithm (FS) resulted 90.60% accuracy. This process can help educational institutions to identify the problems that may minimizes student`s performance and also discusses the effect of combining feature selection with classification on the environment of cloud computing[8].

2. Related Work

In the field of data mining and the analysis of educational activities, many studies have been presented, Tarik et al. [9] discussed the Moroccan students' performance in order to lead the educational process during covid-19 pandemic.

Al fairouz and Al-Hagery [10] used Business and Economics college information to compare the classification results of three models, they got good accuracy to help raising the academic grades of their students and minimizing their weaknesses. Atlam et al.[11] Analyzed the covid-19 impact on psychological health of university students using data from different countries feed it into machine learning models they concluded that the lock down that came during the pandemic harmed the educational process and they suggested some solutions to enhance the online teaching.

Akour et al. [12] proposed a planned behavior theory to study the role of mobile learning platforms in student learning process; they collected student's information from several universities they found that J48 classifier gave best prediction result. Also Mirahmadizad et al. [13] studied student's fillings during the pandemic and how much they effected emotionally during the closure of their schools and universities. By dividing the result to positive and negative emotions, they found that students still have agitation about their schools despite the disease conditions. Morchid N. [14] also took student susceptibility in consideration as a new challenge raised .

Llieva et al. [15] proposed a study to highlight the impact of on-line learning on higher education in some universities around the world by using multi-criteria decision-making and machine learning techniques. Aiming to understand the diminutions of alertness happened due to using distance learning and presents suggestions to manage the future of teaching with less failure. Also in the field of higher education, Gonzalez et al. [16] used students information from three universities in Spain to analyze the effect of lock down and fined that it has a positive impact on students' performance in different subjects without taking into account cheating answers of student in the online exams. In Abdelkder et al. [17] student satisfaction level (SSL) by using feature selection to evaluate a classification model with the relevant features.

Garris and Fleck [18] discussed how higher education in America changed to online at spring 2020, by asking 482 student about different measure to evaluate the learning process. Bansal et al. [19] used both machine learning and deep learning to analyze an estimated student performance for 15 courses. Spinelli and Pellino [20] discussed how students in India faced the same challenges during this period. In Maiti et al. [21] paper an Augmented Reality blended with instructional strategy to build a virtual application used to study student ability to learn in virtual classroom. Bailey et al. [22] Studied the learning gaps between the two periods before and after Spring 2020 specially for elementary schools in U.S. Kanetaki et al. [23] Greece Western University- Mechanical engineering department the data of first semester was used to study the learning outcomes including the statistical analysis.

Mangshor et al. [24] included 420 secondary students data to evaluate quality of online learning through analyze student learning habits.

3. PROPOSED METHOD

The proposed system showed in fig.1 used machine-learning algorithms to study students' performance and to analyze the main factors that can effect student’s grades, these factors could be the key to understand students challenges and to help enhancing their academic levels.

Number of steps made to predict students’ performance:

1. We collect our data from College of Science/Computer department in TIKRIT University.
2. Build a data set with 23 features and 252 instances in CSV format with two types of data (numerical, nominal)
3. Implement a preprocessing step to reduce the imbalanced instances.
4. Spilt the set of features into four subset (personal, study, family, online activities).
5. Apply classification models (J48,SimpleLogisti, SVM) with each subset of features.
6. Compare the performance of each subset to find the best results in Accuracy, ROC, and Recall.

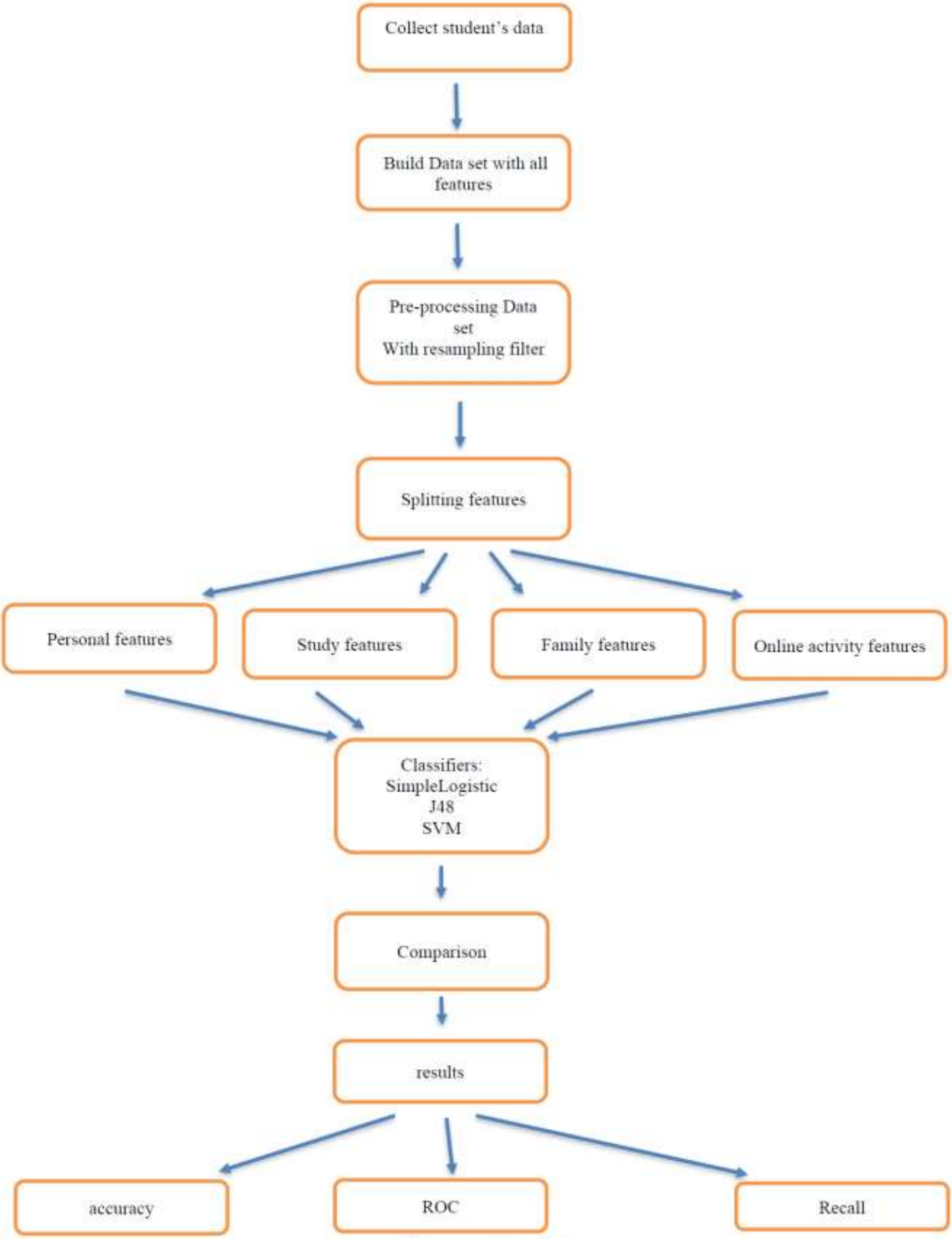

Fig1 proposed method

3.1. Collect Data

Our data collected from Computer Science Department in TIKRIT University, Iraq for two courses in 2021-2022. The collected data considered to:

1- Has an impact on student`s grades.
2- Provides insight into the student's concerns.
3- Students' grades validated.

3.2. Build Dataset

The collected data used to build a (23 attributes, 252 instance) dataset and name class: as target label which is the student's grade.

3.3. Preprocessing

After building the dataset, a preprocessing step applied to enhance the quality of data. The preprocessing methods considered important in data mining algorithms since it involves deleting the duplicated records, filling the missed values, selecting attributes and filtering data. Before classifying data, the preprocessing step can make the data readable by the models and can adjust the imbalance features or instances.

In this step, a Resampling filter applied on instances to improve data balancing.

The Resampling filter: produces a random subsample of the dataset with either replacement or without replacement.

3.4. Splitting Features

In order to measure the influence of data attributes and to conclude which category has the greatest impact on students grades, the attributes divided into four categories as shown in Table 1.

Table 1 The Attributes subsets

Category of attributes	Sub sets of attributes	Description
Personal	ID	Students ID
	Gender	Students Gender
	Marriage	Student marital status
	Work	Student has a job or not
	Special- Sch	Student enrollments in private school
	Time	Number of hours student need to arrive college
Study	Level	stage of students
	Study type	Student study type in college
	Study- Re	Reason to choose the college
	No. failures	Number of failure years in college
	Study -hours	Number of study hours a day
	Academic High-study	Student intention into higher education
	Missed-lectures	Number of missed lectures during the course
Family	Family-size	Student family size
	Mother-Dg	Students mother education degree
	Father-Dg	Students father education degree
	Mother-Jb	Students mother job
	Father-Jb	Students father job
Online Activity	Good-internet	Is the service good at student network
	Online-study	Number of online study hours
	Social media-h	Number of hours student spent on social media
Class: Student final grade (class label in the dataset)		

3.5 Classification Models

This study measures the impact of four types of attributes on students' performance dataset by applying three types of classifiers J48, SVM, and SimpleLogistic. Each category examined by the three models to predict labels of students' grades as shown in fig1.

4. Results And Discussion

This section discusses the result obtained from three models Decision tree J48, SVM, and SimpleLogistic . After applying each model with the selected set of attributes three measures calculated Accuracy, ROC, and Recall.

4.1 Evaluation of the four attributes categories:

- "Personal" attributes evaluated by three measurements in the predictions of three classifiers and noticed that J48 achieved highest results in at accuracy, ROC, and Recall. As shown in Table 2.
- "Study" attributes evaluated by three measurements in the predictions of three classifiers and noticed that J48 achieved highest results in at accuracy, ROC, and Recall. SVM and SimpleLogistic had very close results as shown in Table 3.
- "Family" attributes evaluated by three measurements in the predictions of three classifiers and noticed that SVM achieved highest results in at accuracy, ROC, and Recall results. As shown in Table 4.
- "Online-activity" attributes evaluated by three measurements in the predictions of three classifiers and noticed that J48 achieved highest results in accuracy, and Recall. However, the best ROC result obtained from SimleLogistic. As shown in Table 5.

Table 2: Evaluating "Personal" Attributes

Model	Accuracy	ROC	recall
SimpleLogistic	62.302	0.544	0.623
J48	66.667	0.761	0.667
SVM	62.616	0.524	o.623

Table 3: Evaluating "Study" Attributes

Model	Accuracy	ROC	recall
SimpleLogistic	64.683	0.744	0.647
J48	67.857	0.782	0.679
SVM	64.683	0.610	0.647

Table 4: Evaluating "Family" Attributes

Model	Accuracy	ROC	recall
SimpleLogistic	58.730	0.537	0.587
J48	63.0952	0.528	0.631
SVM	64.683	0.603	0.647

Table 5: Evaluating "Online-activity" attributes

Model	Accuracy	ROC	recall
SimpleLogistic	62.302	0.570	0.623
J48	64.683	0.557	0.647
SVM	62.302	0.526	0.623

4.2 Confusion Matrices Comparison

Although for each subset of attributes, the confusion matrices configured to compare the performance of the classifiers in the manner of class distribution. Noticed the following:

1- When the personal attributes applied, we noticed that J48 produced the best class distribution; SVM and SimpleLogistic make predictions for one class "B" as shown in Table 6.

2- When the "study" attributes selected, we noticed better distribution in prediction labels than " personal" attributes as shown in Table 7..

3- When the " Family" attributes selected, we noticed that all classifiers predict labels for class "B" only, as shown in Table 8.

4- When the " Online-activity" attributes selected, we noticed that J48 produced better class distribution science it predict for two classes "B" and "D"; SVM and SimpleLogistic make predictions for class "B" only as shown in Table 9.

Table 6: Confusion matrices of "Personal" attributes:

SimpleLogistic						J48						SVM						
a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	
0	22	0	0	0	0	5	12	2	3	0	1	0	23	0	0	0	0	a=A
0	157	0	0	0	0	5	140	4	4	3	1	0	157	0	0	0	0	b=F
0	19	0	0	0	0	3	10	1	0	4	1	0	19	0	0	0	0	c=C
0	30	0	0	0	0	4	7	1	15	2	1	0	30	0	0	0	0	d=D
0	18	0	0	0	0	2	7	0	3	6	0	0	18	0	0	0	0	e=B
0	5	0	0	0	0	1	2	0	0	1	1	0	5	0	0	0	0	f=E

Table 7: Confusion matrices of "Study" attributes

SimpleLogistic						J48						SVM						
a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	
3	14	0	4	0	2	9	10	0	2	1	1	4	15	0	4	0	0	a=A
0	152	1	3	1	0	3	140	3	6	5	0	1	153	0	2	1	0	b=F
0	16	0	2	1	0	2	11	3	3	0	0	0	18	0	1	0	0	c=C
2	20	0	8	0	0	3	13	0	13	1	0	2	22	0	6	0	0	d=D
1	14	0	3	0	0	1	9	0	2	5	1	1	16	0	1	0	0	e=B
0	1	0	4	0	0	1	1	0	1	1	1	2	1	0	2	0	0	f=E

Table 8: Confusion matrices of "Family" attributes

SimpleLogistic						J48						SVM						
a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	
0	27	0	0	0	0	2	21	0	0	0	0	6	17	0	0	0	0	a=A
1	148	0	0	0	0	0	157	0	0	0	0	0	157	0	0	0	0	b=F
1	18	0	0	0	0	0	19	0	0	0	0	0	19	0	0	0	0	c=C
0	29	0	0	0	0	0	30	0	0	0	0	0	30	0	0	0	0	d=D
0	19	0	0	0	0	0	18	0	0	0	0	0	18	0	0	0	0	e=B
0	9	0	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	f=E

Table 9: Confusion matrices of "Online-activity" attributes

SimpleLogistic						J48						SVM						
a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	
0	23	0	0	0	0	0	19	0	4	0	0	0	23	0	0	0	0	a=A
0	157	0	0	0	0	0	157	0	0	0	0	0	157	0	0	0	0	b=F
0	19	0	0	0	0	0	17	0	2	0	0	0	19	0	0	0	0	c=C
0	30	0	0	0	0	0	24	0	6	0	0	0	30	0	0	0	0	d=D
0	18	0	0	0	0	0	18	0	0	0	0	0	18	0	0	0	0	e=B
0	5	0	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	f=E

4.3 Measurements Comparison

The dataset classified with three types of classifiers to compare the results and ensure their validity,

The comparison will be between:

1- The accuracy of three classifiers for each subset of attributes. As shown in Table 10 and Fig2.
   - "Study" attributes achieved best accuracy result and the higher result for J48 classifier.
   - In the second level "Personal" then "Online-activity", and the less accuracy for "family" attributes for the SimpleLogistic model.
2- The ROC of three classifiers for each subset of attributes. As shown in Table 11 and Fig3.
   - "Study" attributes achieved best ROC result and the higher result for J48 classifier.
   - In the second level, "Personal" then "Online-activity" and the less ROC for "family" attributes because it has the less ROC result for SimpleLogistic model.
3- The Recall of three classifiers for each subset of attributes. As shown in Table 12 Fig4.
   - "Study" attributes achieved best Recall result and the higher result for J48 classifier.
   - In the second level, "Personal" then "Online-activity" and the less Recall for "family" attributes because it has the less Recall result for SimpleLogistic model.
4- Best obtained accuracy for the four subsets. As shown in Table 13.
   - "Study" attributes resulted highest accuracy by using the tree classifier J48.
   - All the subsets obtained best accuracy by using J48 except for "Family" subset achieved best accuracy with SVM model.

Table 10: Accuracy Comparison

Model	"Personal" Accuracy	"Study" Accuracy	"Family" Accuracy	"Online" Accuracy
SimpleLogistic	62.302	64.683	58.730	62.302
J48	66.667	67.857	63.0952	64.683
SVM	62.616	64.683	64.683	62.302

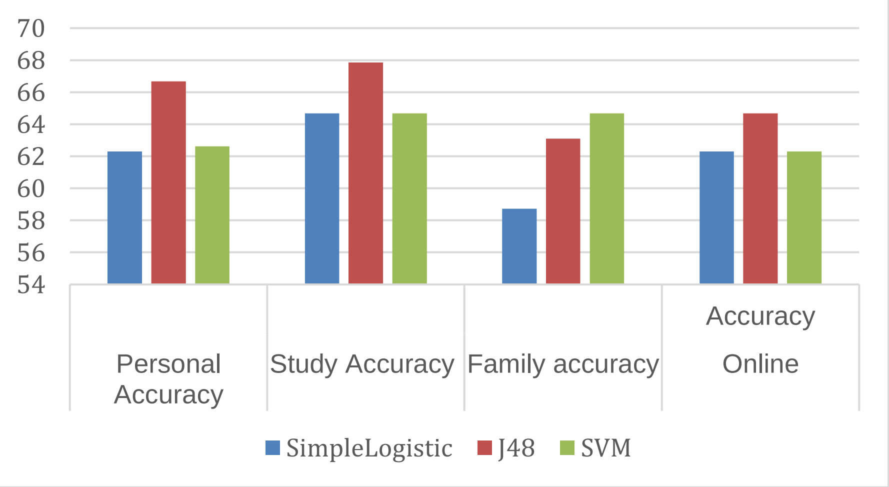

Figure 2: Accuracy comparison

Table 11: ROC comparison

Model	"Personal" ROC	"Study" ROC	"Family" ROC	"Online Activities" ROC
SimpleLogistic	0.544	0.744	0.537	0.57
J48	0.761	0.782	0.528	0.557
SVM	0.524	0.61	0.603	0.526

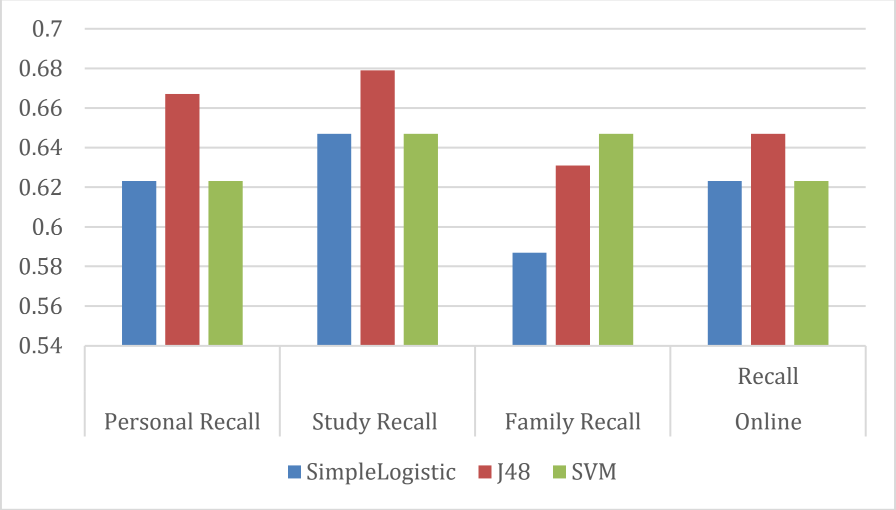

Figure 3: ROC comparison

Table 12: Recall comparison

Model	Personal Recall	Study Recall	Family Recall	Online Recall
SimpleLogistic	0.623	0.647	0.587	0.623
J48	0.667	0.679	0.631	0.647
SVM	o.623	0.647	0.647	0.623

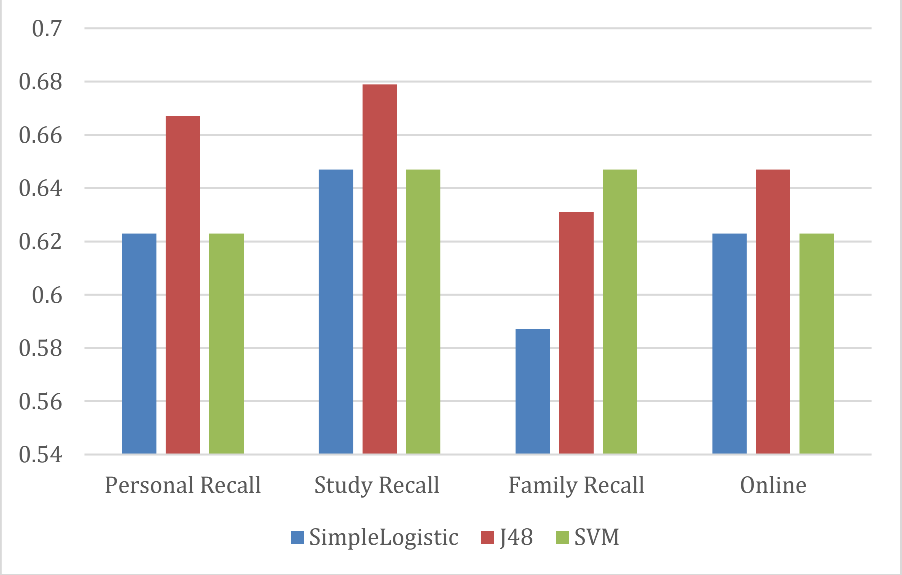

Figure 4: Recall comparison

Table 13: Compare best accuracy

Attribute Category	Best accuracy	classifier
Personal	66.667	J48
Study	67.857	J48
Family	64.683	SVM
online	64.683	J48

5. Conclusions

Analyzing student performance helps our educational institutions to obtain knowledge that can enhance the quality of learning. Data mining techniques offers the ability of analyzing student's data and predicting their outcomes, although these outcomes effected by some circumstance. In this study we labeled the attributes of our students in four categories (personal, study, family, and online activity) to predict which category has the most impact on the students' grade. "Study" attributes showed a great impact on students' performance since the three classifiers produce best results in accuracy, ROC and Recall measures. Best accuracy achieved by using "Study" attribute with J48 model 67.857. Also in the comparison of models, the classifier J48 achieved best results and more relevant predictions.

References

1. B. K. Francis and S. S. Babu, "Predicting Academic performance of Students Using a Hybrid Data Mining Approach", Journal of medical systems, 43;162, 2019.
2. M. N. Ismael,"Students Performance Prediction by Using Data Mining Algorithm Techniques", vol. 6, ISSN:2795-7640, 2022.
3. W. Punlumjeak and N. Rachburee, "A Comparative Study of feature Selection Techniques for Classify Student Performance", International conference on Information technology and electrical engineering, 2015.
4. R. Bertolini, S. J. Finch, and R. H. Nehm, "Enhancing Data Pipelines for Forecasting Student performance; Integration Feature Selection with Cross-Validation", Internaional Journal of education technology in higher education 18, 44(2021).
5. M. Zaffar, M. A. Hashmani, K. S. Savita, and S. S. Hussain Rizvi, "A Study of Feature Selection Algorithms for Predicting Students Acadimic Performance", vol. 9, no. 5, 2018.
6. P. Sokkey and T. Okazaki, "Student on Dominant Factor for Academic Performance Prediction Using Feature Selection Methods", vol. 11, no. 8, 2020.
7. Md. A. Arif, A. Jahan, M. I. Mau, and R. Tummarzia, "An Improved Prediction System of Students` Performnce Using Classification Model and Feature Selection Algorithm", vol. 13, no. 1, 2021.
8. W. Punlumjeak, N. Rachbaree, and J. Arunrerk, "Big Data Analytics: Student Performance Prediction Using Feature Selection and Machine Learning on Microsoft Azure Platform", vol. 9, no. 1-4, 2017.
9. A.Tarik, H.Aissa and F. Yousef, " Artificial Intelligence and Machine Learning to Predict Student Performance During the COVID-19", Proceeding computer science , pp. 835-840, 2021.
10. E. Al fairouz and M. Al-Hagery, " The Most Efficient Classifiers for the Students` Academic Dataset", vol. 11, No. 9, 2020.
11. E. Atlam, A. Ewis, M. El-Raouf, O. Ghoneim and I. Gad, "A new approach in identifying the psychological impact of COVID-19 on university student's academic performance", Alexandria Engeneering Journal, vol. 61, pp. 5223-5233, 2022.
12. I. Akour, M. Alshurdan, B. Kardi, A. Ali and S. Salloum, "Using Machine Learning Algorithms to Predict People's Intention to Use Mobile Learning Platforms During the COVID-19 Pandemic: Machine

Learning Approach", JMIR Med Idue, vol. 7, issue 1, p. 1, 2021.

13. A. Mirahmadizad, K. Ranjbar, R. Shahriarirad, A. Erfani and T. Rohimi, "Evaluation of students' attitude and emotions towards the sudden closure", vol. 8, 2020.
    of schools during the COVID-19 pandemic: a cross-sectional study",
14. Morchid N., " The Current State of Technology Acceptance: A comparative Study", vol. 22, Issue 2, pp 01-16, 2020.
15. G. Llieva, T. Yankova, S. Klisarova-Belchera and S. Ivanova" Effects of COVID-19 Pandemic on University Students' Learning", vol. 12, 2021.
16. T. Gonzalez, M. Delapubia, K. Hincz and S. Fort, "Influence of COVID-19 confinement on students' performance in higher education", PLOS One , 15(10), 2020.
17. H. Abdelkder, A. Gad and S. Sarour"An Efficient Data Mining Technique for Assessing Satisfaction Level With Online Learning for Higher Education Students During the COVID-19", vol. 10, 2022.
18. C. Garris and B. Fleck, "Student Evaluations of Transitioned-Online Courses During the COVID-19 Pandemic", vol. 8, No. 2, pp 119-139, 2022.
19. V. Bansal, H. Buckchash and B. Rman, "computational Intelligence Enabled Student Performance Estimation in the Age of COVID-19", SN Computer Science 3, No. 41, 2022.
20. A. Spinelli and G. Pellino, " COVID-19 Pandemic: Perspectives on an Unfolding Crisis", 107(7): 785-787, 2020.
21. M. Maiti, M. Priyaadharshini and B. Sundaram, "Augmented Reality in Virtual Classroom for High Education During COVID-19", Intelligent Computing, vol. 285, pp 399-418, 2021.
22. D. Bailey, G. Duncan, R. Murnane and N. Yeung, "Achievement gaps in the wake of covid-19", vol. 50, Issue 5, 2021.
23. Z. Kanetaki, C. Stergiou, G. Bekas, C. Troussas and C. Sgouropoulou, " Analysis of Engineering Student Data in Online Higher Education During the COVID-19 Pandemic", iJEP, vol. 11, No. 6, 2021.
24. N. Mangshor, S. Ibrahim, N. Sabri and S. Kamaruddin, " Students` Learning Habit Factors During COVID-19 Pandemic using Multilayer Perceptron (MLP)", vol. 8(74), ISSN (print): 2394-5443, 2021.