# An Intelligent Approach for the Effect of Social Media on Undergraduate StudentsṔerformance: A Case Study in the University of Jordan


Yazn Alshamaila
The University of Jordan
Amman, Jordan
y.shamaileh@ju.edu.jo

Maria Habib
The University of Jordan
Amman, Jordan
neutrinoscientist@gmail.com

Ibrahim Aljarah
The University of Jordan
Amman, Jordan
i.aljarah@ju.edu.jo

Hamad Alsawalqah
The University of Jordan
Amman, Jordan
hamad.alsawalqah@gmail.com

Hossam Faris
The University of Jordan
Amman, Jordan
hossam.faris@ju.edu.jo

Anas AlSoud
Al-Ahliyya Amman University
Amman, Jordan
a.alsoud@ammanu.edu.jo



## ABSTRACT

Promoting the learning environment within Jordanian universities and maximizing the students' academic gain are essential national-level problems. Since learning and teaching systems are main building bricks to grow individuals who are responsible for developing and flourishing culture and civilization for Jordanian society. Educational data mining focuses on developing new smart algorithms devoted to analyzing the resulted data from educational systems; in order to better understand students and the learning environments. In this paper, we are analyzing major factors affecting university students' performance and the effect of social media usage on them. Furthermore, predicting the students' performance by adopting different rule-based data mining algorithms like rule learner based on Repeated Incremental Pruning to Produce Error Reduction (JRIP) and a type of decision tree called (PART). We have conducted a research survey within the University of Jordan students that is covering all faculties and cover a vast range of different students. Using both JRIP and PART we have concluded fundamental remarks; mainly, we have noticed that using YouTube as a learning resource has positive impacts on students' performance especially within scientific faculties. Moreover, we have interpreted the impact of other factors, such as having an Internet connection, having several social media applications and others. Certainly, upon our findings, we recommend the importance of integrating YouTube as a learning resource within universities learning environments.



## CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; • **Applied computing** → *Education.*




*ICCIT 2020, April 14–16, 2020, Antalya, Turkey*



ACM ISBN 978-1-4503-7749-2/20/04…$15.00
https://doi.org/10.1145/3397125.3397146



## KEYWORDS

Machine Learning, Educational Data Mining, Social Media, University of Jordan


## 1 INTRODUCTION

In the new educational environment of universities, which is highly complex and competitive, the students' performances have become a central issue for pointing out the uniqueness of the educational institution, as well as building strategies for future developments. In the last couple of decades Jordanian universities are competing not only in a national level but on an international level as well, thus according to the Quacquarelli Symonds (QS) ranking the University of Jordan is ranked between number 551-600 internationally, number nine in the Arab region and number one nationally. Thus, one of the greatest challenges for these universities' management is to use the admitted students' profiles and their entry points as resources to help build new marketing campaigns, build strategies to increase the performance of weak students, in addition, to encourage the potential of promising students.

The study will be conducted on collected data from an online survey distributed among university students. The collected data will be analyzed using data mining techniques[1–3]. The implementation of data mining techniques in acquiring new patterns and information from the collected data is the central focus of this research study. In this research, we are trying to prove the high potential of data mining implementation in educational management generally and in universities specifically.

The educational system in Jordan covers different levels from preliminary studies until graduate and post-graduate studies. the higher education system was introduced in Jordan in 1951 and has been on a continuous development since then [4]. The first public university established in Jordan is the University of Jordan (UJ) in 1962, which is the data source for this study. According to the statistics released by the Ministry of Higher Education in Jordan (MOH), over a decade almost 7500 students graduate annually from UJ alone, in which 2000 students are male and the remaining 5000 are female students [5].

Educational data mining (EDM) [6, 7] is a new discipline that focuses on developing new smart methods and algorithms devoted to analyzing the data that resulted from educational systems; in order to better understand the students and the learning environment. Data mining is also known as knowledge discovery in databases (KDD), which is the process of discovering novel valuable information from a large amount of data stored in databases or data warehouses [8]. In this paper, we focus on class of data mining that concerned with rule-based machine learning techniques. Rule-based learning techniques apply such types of learning algorithms; in order to produce and identify useful rules [9]. Those rules take the form of if-then rules which is relatively easier for people to read and understand [10]. However, there are different approaches for rule-based machine learning as; rule-learning classifiers systems, association rule mining, and artificial immune systems [11]. All of them rely on extracting rules to express contextual descriptive knowledge [11]. In this paper, we focus on rule-based learning classifiers. Rule classifiers aim to identify a set of rules that represent accurate classifiers [11].

In this paper, we investigate the use of social media in the learning environment and examine the possible positive and negative effects of using social networking applications and sites on student academic performance. As a consequence, we identify the major factors which play an important role in the students' performance as well as in predicting students' performance based on several questions displayed to the students as a survey.

This first section 1 introduces the topic and the educational system in Jordan in addition to explaining the importance of this paper to the local universities in improving the managerial strategies. The remaining parts of the paper proceed as follows. Section 2 will display previous studies, applications of data mining implementation, and its use in educational fields. In addition, a background. Section 3 explains the methodology used in the analysis process. While section 4 discusses the results and findings. Finally, section 5 is the paper ends with a summary conclusion and possible future works.

# 2 BACKGROUND AND RELATED WORKS

## 2.1 Rule-Based Classifiers

Rule-based classifiers are rule-learning classifiers that use a set of if-then rules to do classification. If-then rules let us represent the inferred knowledge of a dataset in an easy way for people to understand, which is one of the merits of rule learning classifiers over decision tree learner classifiers. Nonetheless, since decision trees suffer from overfitting, reduced error pruning (REP) technique is used in rule learning classifiers to overcome this problem [12]. In this study, we utilize both JRip and PART algorithms to analyze the students' academic performance.

*2.1.1 JRIP Algorithm.* JRip is the Weka workbench implementation of a propositional rule learner based on Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Created by William W. Cohen. as an enhanced version of IREP. Mainly, JRip goes through two stages; the building stage and the optimization stage. The building stage iteratively enters two phases; a growing phase then a pruning phase [9].

The growing phase grows one rule by a time. Starting by an empty rule and greedily adding conditions to the rule until it is 100% accurate. In this phase all possible conditions of all attributes will be tried, thus, the algorithm picks the condition with the highest information gain, as given by Eq. 1 [9].

$$IG(S,F) = H(S) - \sum_{f \in F} \frac{|S_f|}{|S|} * H(S_f) \tag{1}$$

In which S is a set of features. $S_f$ is the number of elements of S having feature F with value f. Whereas, H(S) is the entropy of S and given by Eq. 2.

$$H(S) = -\sum_{c \in C} p_c * log_2\, p_c \tag{2}$$

$P_c$ is the probability of elements of S belonging to class c. While S is a set of examples with C classes.

The pruning phase, incrementally prunes each rule by deleting any final sequence of conditions that maximize the pruning metric $v$ given by Eq. 3 [9].

$$v = \frac{p - n}{p + n} \tag{3}$$

$n$ is the number of negative examples covered by the rule in the validation set. $p$ is the number of positive examples covered by the rule in the validation set.

The building step keeps repeating the growing and pruning phases until there are no positive examples left, the description length value of a rule is greater than a predefined threshold value, or the error rate is >= 50% [9]. Upon finishing the building stage; an initial rule set is produced to enter the optimization step.

The optimization stage takes the produced rule set and creates two variants of each rule. Then optimize the new rules by again using the growing and pruning steps, in addition to a modified version of the pruning metric, given by Eq. 4 [9].

$$v = \frac{TP + TN}{P + N} \tag{4}$$

TP is same as p, and TN is same as n. P and N is the total number of positive and negative examples, respectively [9].

The rule with minimal description logic is selected as the final representative rule [9]. As the final step, the algorithm deletes any rule from the rule-set that would maximize the description length of the whole rule-set.

*2.1.2 PART Algorithm.* Primarily, it is called PART because it is based on partial decision trees. PART works by using a separate-and-conquer strategy to generate a decision list. It builds a partial C4.5 decision tree in each iteration and converts the best leaf node into a rule [10]. C4.5 algorithm builds decision trees from the training set based on the information entropy. The attribute with the highest information gain selected to split the examples into subsets. Then the algorithm keeps recursing on the produced subsets [10].

## 2.2 Related Works

Recently, several technologies for collecting and generating new data have been developed. Data mining can be defined as the process of finding meaningful new patterns and correlation by mining into a tremendous amount of data using certain techniques such as statistical, and machine learning [13]. This technology is not related to one industry; it has been used in several application fields like medical, industrial manufacturing, and many more. The integration of data mining techniques in decision support systems; promotes the decision-making process and help in building more reliable trustworthy systems [13].

Data mining can be viewed as an evolution of information technology [14]. EDM refers to the utilization of data mining tools in analyzing the existing data in the educational systems and institutions [15]. Authors in [15] conducted a survey in order to investigate the integration of data mining in education from 1995 to 2005. They found that a great deal of previous research has focused on educational data mining. Educational data mining helps researchers in targeting a variety of questions related to the psychology of learning [15]. Cognition, language, motivation, and social discourse are some of the issues addressed by implementing educational data mining using data resulted from intelligent systems, online courses, and discussion forums [16].

The application of EDM covers areas such as educational software, computer-adaptive testing and the factors which impact the students' performance [8]. Educational data mining is very useful in identifying the priority needs of the students and assessing the institutional performance [17]. In higher education institutions, Educational data mining is used to concentrate on improving institutional effectiveness, enrollment management and retention of students [18].

Factors that influence the performance of the students have been viewed in the literature. In [19] the authors conducted a review to show data mining techniques used on the prediction of the students' performance. In their systematic review, they summarize the factors which affects the students' performance as cumulative grade point average, internal assessment such as assignment mark, quizzes and lab work, demographic factors including gender, age, family background, and disability, external assessments, extra-curricular activities, high school background, social interaction network and psychometric factor which includes student interest and family support. Gender has been found as one of the significant factors which influence the performance in several studies [20], [21], [22]. In those studies, females perform better than males. Authors in [20] justifies that most females have positive learning styles and behaviors in comparison with males. Female students are more focused, self-directed, and perform well since they have effective and successful learning strategies [21]. Cumulative grade point average (CGPA) is found as a crucial and the most important factor in [23], [24], [25]; as CGPA has a real value for future educational mobility [19].

There are several features that affect the students' performance such as gender, income, smartphones and social media. In [26] the researchers stated that the grade point average for females is higher than males' GPA. Other research like [27].

Predicting and analyzing students' performance is a hot research topic, since it is an essential argument for educational systems, either locally or globally. The education operation is complex and heterogeneous among countries. However, in [28] authors studied the school systems of nine developed countries and found that some of the most influential factors on students' performance are socio-economic index, anxiety, motivation, gender, and parental education [28]. Nonetheless, in [29] authors claim that receiving financial support decreases the students' performance, moreover, they state that race, gender, and past academic history significantly affect students' grade point average (GPA). Nonetheless, in [30] authors identified a set of factors affecting undergraduate Latin students' performance, such as socio-cultural features, university experiences, and communication with helpful individuals.

Nowadays, with the continuous evolution of technology; we see a tremendous trend in using the evolved digital computing devices, and digital social applications. In [31] 67% of the surveyed students believe that the use of mobile devices has an important role in their success and in their academic activities' usage. However, [32] argued the role of social media in higher education platforms, in which they discussed existing limitations for deploying social media as an educational tool. Nonetheless, in [33] authors found that the adoption of Facebook as an educational instrument for higher education, had different benefits as increased student to teacher and student to student interactions, also improved performance. Even so, the debate on the benefits and usage of Facebook in educational platforms continues [33]. Typically, the use of visual teaching techniques has a fundamental effect on learner [34]. In [34] the researchers studied the impact of integrating YouTube as a complementary teaching tool. They did a study on different university students. As a result, they encourage instructors to adopt YouTube resources in their activities and course material. Additionally, in [35] the authors conducted a study on UK university students to test the integration of social media in a higher education institutions. They suggest recommendations to consider before the practical adoption of social media in education.

# 3 METHODOLOGY

In this section, we present a detailed explanation of our designed approach and the collected dataset.

CRISP-DM is one of the most used tools in data mining analysis [18], thus we utilize it in our designed approach. It uses a cyclic approach that undergoes six main phases; Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In order to run our experiments, we use WEKA workbench. [36] authors state that WEKA software is easy to use and known for its wide range of classification methods.

## 3.1 Dataset Description

Basically, the data have been collected through an online survey and cloud-based software called SurveyMonkey [37]. The survey launched at the University of Jordan, during summer 2018 within Social Media course, over two months. The Social Media course is an elective course where students from any faculty can attend it. The design of the questionnaire and choosing the types of questions both are key aspects in developing the dataset. The objective of the survey is to collect information or distinctive features of undergraduate students.

## 3.2 Designed Approach

This section discusses our designed methodology procedure as will be mentioned in the following subsections. Figure 3 shows an overview of the designed approach.

*3.2.1 Data Preprocessing.* The first essential step in order to analyze our data is the preprocessing step. The quality of data plays a

**Table 1: Dataset Description**

| Feature | Type | Values | Description |
|---|---|---|---|
| Gender | Nominal | Male, female | Student's gender |
| Age | Numeric | [18, 19, 20, 21, 22, >23] | Student's age |
| Faculty | Nominal | Scientific, humanities, health | Field of study |
| Study year | Numeric | [1, 2, 3, 4, 5] | Current study year |
| Income | Numeric (JD) | [(50-100), (100-150), (150-200), >200] | Student's income |
| Internet availability | Nominal (binary) | Yes, no | Having Internet connection at home |
| Internet Type | Nominal | 3G, wifi, no_Internet | Type of Internet they are using |
| Internet using experience | Numeric (year) | [ < 1 ,(1-2), (3-5), (5-7), > 7] | The experience of using Internet less than year or greater than 7 years |
| Internet using time | Numeric ( hr/ day) | [ 1, (1-2), (2-3), (3-4), (4-5), > 5] | The number of hours a student use the Internet in a day |
| Mobile type | Nominal | Android, apple, other | The manufacturer of the mobile |
| Mobile using time | Numeric (hr/ day ) | [ 1 , (1-2), (2-3), (3-4), (4-5), > 5] | The number of hours a student use his/her mobile in a day |
| Facebook | Nominal (binary) | Yes, no | Having a Facebook account |
| Followers | Numeric | [ <50 , (50-100), (100-300), (300-500), (500-1000), (1000-2000), (2000-3000), >3000] | Number of Facebook followers |
| Twitter | Nominal (binary) | Yes, no | Having a Twitter account |
| Instagram | Nominal (binary) | Yes, no | Having an Instagram account |
| Snapchat | Nominal (binary) | Yes, no | Having a Snapchat account |
| YouTube | Nominal (binary) | Yes, no | Having a YouTube account |
| Linked In | Nominal (binary) | Yes, no | Having a Linked In account |
| Pinterest | Nominal (binary) | Yes, no | Having a Pinterest account |
| Social media using time | Numeric ( hr/day) | [ <0.5 hr/week , <0.5, (1-1.5), (1-2), (2-3), (3-4), >4] | The number of hours a student use social media in a day |
| Average | Nominal | Excellent, very good, good, weak, very weak | The student cumulative academic average |

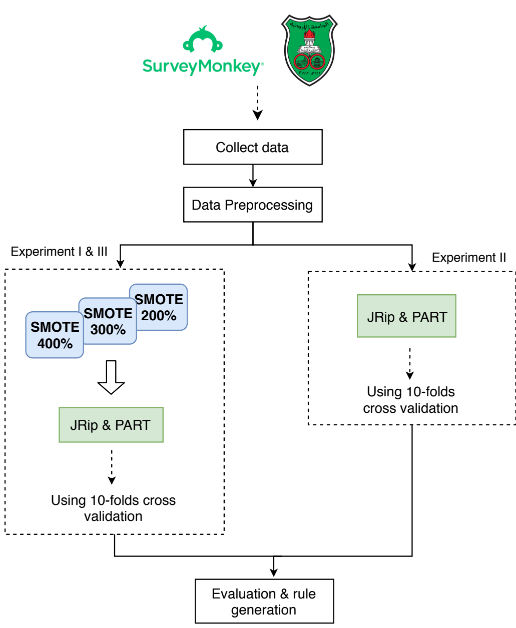


**Figure 1: Designed Approach**

significant role in the reliability of the classifier. Having redundant, noisy, and irrelevant features deteriorate the performance of the learning method [38]. Therefore, by examining our data there are several missing values across different features as well as some data instances do not have a class label. Consequently, we have applied a filter to remove the missing values for specific features with the mode value of that feature. According to the unlabeled class values we have removed the corresponding data instances. We conducted our analysis on the processed data through three experiments which are illustrated as follows.

*3.2.2 Experimental Setup.* All the experiments and the preprocessing step conducted using Weka machine learning workbench [12]. In all the experiments we use both JRip and PART algorithms. The parameter settings for both algorithms are set to their default values and as identified by the Weka library. Except for the number of folds is set to 10. The number of folds specifies the amount of data to be used. In which, one-fold used for the pruning, and the rest used for growing the rules. The datasets are evaluated using 10-folds cross-validation.

*3.2.3 Experiment I.* The first experiment arranges the students into two groups; pass and fail. We consider the *Average* attribute is the class attribute. Since the *Average* class consists of five values. Hence, the excellent, very good, and good averaged students are representing the pass class, while weak and very weak students are representing the fail class. Table 2 shows the classes' percentages over all the three experiments. Obviously, in experiment I the dataset is imbalance and the minor class is the fail class with

**Table 2: Dataset Classes Distribution**

| Experiment | Class | Percentage |
|---|---|---|
| Experiment I | Pass | 84% |
| | Fail | 16% |
| Experiment II | A | 50% |
| | D | 50% |
| Experiment III | A | 15% |
| | B | 69% |
| | D | 16% |

a percentage of 16 % of all the data. It has been proofed empirically that class imbalances decrease the performance of standard classifiers [39]. One of the approaches to deal with the imbalance problem is a filter-based technique called SMOTE; SMOTE is Synthetic Minority Over-sampling Technique. In which the minority class is over-sampled by creating new interpolated instances based on the neighboring instances [40]. Since the data needs to be balanced, we have run SMOTE at three percentages 200%, 300%, and 400%. Afterward, the resultant data applied to both JRip and PART algorithms.

*3.2.4 Experiment II.* In this experiment, we excluded the students who have good and very good averages. Therefore, the dataset contains two classes; the first class is the class that represent the excellent averaged students, pointed out as A. Whereas, the second class represents the weak and very weak averaged students, pointed out as D. Noticeably, both of the extreme classes A, and D are having equal percentages of data instances, thus the dataset is balanced. Next, the data directly applied to both JRip and PART algorithms.

*3.2.5 Experiment III.* In the third experiment, we have categorized the students' averages into three groups. The first group refer to the collection of students who have excellent Average, those are denoted as class A. The second group represents the group of students who have either very good or good average, and denoted by class B. Finally, the third group represents the students who have weak and very weak average, denoted by class D. Obviously, referring to table 2 both class A and D are representing a minority classes that need to be over-sampled. The data in this experiment is over-sampled at three levels when SMOTE percentage is 200%, 300%, and 400%. Thereafter, the three variants of data are applied to both JRip and PART algorithms.

# 4 RESULTS AND DISCUSSION

## 4.1 Evaluation Metrics

In order to evaluate the performance of the algorithms we use different performance metrics as accuracy, recall and G-mean [41–43]. Accuracy: is the ratio of correctly classified students (into for example fail and pass) over the total number classified students (Eq. 5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Recall: known as sensitivity (Eq. 6). Measures the percentage of true positives (TP) that are correctly identified as positives.

$$Recall = \frac{TP}{FN + TP} \tag{6}$$

G-mean: it is the Geometric mean (Eq. 7). It measures the classification reliability using the square root of the product of both the specificity and the sensitivity. In which the specificity is the percentage of true negatives (TN) that are correctly identified as negatives.

$$G - mean = \sqrt{specificity.sensitivity} \tag{7}$$

## 4.2 Discussion and Experiments Results

Table 3, table 4, table 5 show the results of deploying the constructed data from experiment I on both JRip and PART.

It is obvious that both JRip and PART algorithms perform well in terms of accuracy and recall. For example, JRip achieves 81% accuracy and PART achieves 73%. However, the more important is the resulted rules. We can observe that JRip obtained a few numbers of rules (which is 9) in comparison to PART which obtained 116 when SMOTE is 200%. However, we did not notice a valuable increase in performance with increasing the percentage of SMOTE. Therefore, we collected all the rules from both algorithms from the three SMOTE variants and explored the rules for any frequent pattern of data. By analyzing the resulted rules, we come up with various important patterns of information as follows.

We found that students from scientific faculties do not use YouTube. That is affected their academic performance negatively and they classified with fail average.

- (Faculty=Scientific) and (YouTube =No) then (Avg=fail) (510/199)

The numbers beside the rule as (A/B) gives an indication of the accuracy of the rule. Where A is subject to the number of individuals matching this rule, whereas, B is the number of misclassified instances from A value [13].

However, there is no effect of not using YouTube on students from humanities and Literature background, which return to the nature of the study and its requirements. The following rules are examples of the frequent pattern of rules.

- (Faculty=Humanities) and (Pinterest=No) and (Internet-type=wifi) and (Mobile-using-time= 1-2 hrs) and (Instagram=No) and (followers= 200-300 ) and (YouTube =No) then (avg=pass) (45/0.0)

Additionally, if there is no Internet connection with a relatively long time for using the mobile, this is accompanied by a negative performance. As it in the following rule.

- (Faculty=Scientific) and (YouTube =No) and (mobile-using-time= 4-5 hrs) and (Internet-availability=No) then (Avg=fail) (73/12)

Using the social media for long time besides 3G internet availability affect the performance negatively. For example:

- (Faculty=Scientific) and (YouTube = No) and (mobile-type= Android) and (Internet-type=3G) and (social-media-time= 4 hrs or more) then (Avg=fail) (96/9)

Nonetheless, the non-existence of Internet connection, also has a passive impact on the students. The following rule justifies that:

**Table 3: Results of Experiment I using SMOTE 200%**

| | Accuracy (%) | Recall (Pass) | Recall (Fail) | G-mean | #Rules |
|---|---|---|---|---|---|
| JRip | 81.9076 | 0.939 | 0.182 | 0.413 | 9 |
| PART | 73.5806 | 0.838 | 0.191 | 0.400 | 116 |

**Table 4: Results of Experiment I using SMOTE 300%**

| | Accuracy(%) | Recall (Pass) | Recall (Fail) | G-mean | #Rules |
|---|---|---|---|---|---|
| JRip | 81.1506 | 0.939 | 0.134 | 0.355 | 14 |
| PART | 74.7918 | 0.847 | 0.22 | 0.432 | 120 |

**Table 5: Results of Experiment I using SMOTE 400%**

| | Accuracy(%) | Recall (Pass) | Recall (Fail) | G-mean | #Rules |
|---|---|---|---|---|---|
| JRip | 81.5291 | 0.938 | 0.163 | 0.391 | 19 |
| PART | 75.1703 | 0.846 | 0.249 | 0.459 | 127 |

**Table 6: Results of Experiment II**

| | Accuracy(%) | Recall (Pass) | Recall (Fail) | G-mean | #Rules |
|---|---|---|---|---|---|
| JRip | 64.878 | 0.607 | 0.689 | 0.647 | 6 |
| PART | 64.1463 | 0.612 | 0.670 | 0.640 | 43 |

- (Faculty=Scientific) and (Internet-availability= No) and (mobile-using-time= 4-5 hrs) and (social-media-time ≥ 1.5 hr) and (Snapchat= No) and (Instagram= No) and (Internet-experience ≥ 1-2 yr) then (Avg=fail) (53/1.0)

Table 6 shows the results of deploying the constructed data from experiment II on both JRip and PART. As it clear in table 6, the dataset is balanced; hence, we do not apply SMOTE. Considering the accuracy and g-mean both JRip and PART are doing well and relatively close to each other. For instance, according to the g-mean both JRip and PART achieved 0.647 and 0.640, respectively. By exploring the rules, we found that the resulted rules emphasize, too, on the importance of having Internet connections on students' academic performance. Moreover, and in contrast, having high internet using time and high social media time is accompanied by positive effects on students' performance in this experiment.

Table 7, table 8, table 9 show the results of the deploying the constructed data from experiment III on both JRip and PART. We can see that the JRip algorithm performs better than PART in terms of accuracy and g-mean. However, by anticipating the resulted rules from both algorithms within the variants we can notice that having a good internet experience can increase positively the academic performance of students. As will be shown in the following rules:

- (Faculty=Health) and (Internet-experience ≥ 7 yrs) and (Mobile-using-time≤ 3-4 hrs) then (avg= A) (103/10)
- (Faculty= Humanities) and (Snapchat= No) and (Internet-experience ≥ 7 yrs) and (Age ≤ 21) then (avg= A) (62/9.0)

**Table 7: Results of Experiment III using SMOTE 200%**

| | Accuracy(%) | Recall (A) | Recall (B) | Recall (D) | G-mean | #Rules |
|---|---|---|---|---|---|---|
| JRip | 60.106 | 0.269 | 0.762 | 0.220 | 0.212 | 17 |
| PART | 51.779 | 0.294 | 0.637 | 0.215 | 0.201 | 223 |

**Table 8: Results of Experiment III using SMOTE 300%**

| | Accuracy(%) | Recall (A) | Recall (B) | Recall (D) | G-mean | #Rules |
|---|---|---|---|---|---|---|
| JRip | 55.7154 | 0.323 | 0.686 | 0.220 | 0.221 | 24 |
| PART | 52.3089 | 0.259 | 0.643 | 0.254 | 0.206 | 234 |

**Table 9: Results of Experiment III using SMOTE 400%**

| | Accuracy(%) | Recall (A) | Recall (B) | Recall (D) | G-mean | #Rules |
|---|---|---|---|---|---|---|
| JRip | 50.6435 | 0.299 | 0.605 | 0.278 | 0.224 | 23 |
| PART | 52.7631 | 0.313 | 0.639 | 0.249 | 0.223 | 235 |

- (Faculty= Health) and (Internet-experience ≥ 7 yrs) then (avg= A) (256/33)

Also, it can be noticed that there is sometimes high Internet using time and social media time which accompanied with low academic performance, as follows:

- (Faculty=Scientific) and (YouTube = No) and (Internet-type=3G) and (Internet-using-time= ≥ 4-5 hrs) then (avg= D) (216/38)
- (Faculty=Scientific) and (Internet-using-time= ≤ 4-5 hrs) and (social-media-time ≥ 3-4 hrs) then (avg= D) (103/20)

Furthermore, it can be investigated that students from health faculties who relatively have high income and do not use YouTube; gain excellent average and do well.

- (Faculty=Health) and (YouTube = No) and (Income ≥ 200) and (Age ≤ 20) then (avg= A) (60/3.0)

Further, we observe potential relation with the number of followers, where the students with a relatively high number of followers have high positive academic performance.

- (followers ≥ 300) and (YouTube = Yes) and (Instagram= No) and (Age ≤ 20) then (avg= A) (39/6)
- (followers ≥ 300) and (Internet-type= wifi) and (Sex= female) and (Snapchat= No) and (Age ≤ 21) and (Internet-experience ≥ 3-5 yrs) then (avg= A) (41/2)

## 4.3 Limitations and Future Work

Generally, from the resulted rules we can observe the importance of using the Internet and social media within the learning environments. However, some findings still need more clarification and investigation. Such as the long periods of using the Internet or social media, or the mobile device, whereas, he/she has negative academic performance. This needs a more customized survey that looks for more details on why the use the corresponding technologies. Moreover, we would like to investigate other possible factors of media and social networking that might affect students' performance.

# 5 CONCLUSION

Recently, we are witnessing an increasing development in technology and other digital appliances like smartphones and iPads. In order to adapt to the evolution of information technology; this requires the preparation for more smart environments to elevate the learning process. Typically, the emergent of data mining techniques facilitates the process of analyzing students' behaviors in order to understand the effect of such factors especially the effect of social media on students' academic performance. In this paper, we conducted a survey within the University of Jordan questioning students' habits and attributes on the use of social media and smart applications. All to examine the effect of emergent digital tools on the academic performance of undergraduate students. We utilized two dominant algorithms JRip and PART for extracting such an informative pattern of data. We found that integrating YouTube as a complementary educational tool within course materials is very important for the success of students, especially within scientific faculties and health faculties. Also, the adoption of networking sites as Facebook is a significant advantage to promote active and interactive teaching and learning styles. Nonetheless, having a stable Internet connection is important for the success of students.

# ACKNOWLEDGEMENTS

This research is funded by the Deanship of Scientific Research in the University of Jordan, Amman, Jordan.

# REFERENCES


[1] Hossam Faris, Ibrahim Aljarah, and Ja'far Alqatawna. Optimizing feedforward neural networks using krill herd algorithm for e-mail spam detection. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5. IEEE, 2015.

[2] Hossam Faris, Ibrahim Aljarah, Nailah Al-Madi, and Seyedali Mirjalili. Optimizing the learning process of feedforward neural networks using lightning search algorithm. *International Journal on Artificial Intelligence Tools*, 25(06):1650033, 2016.

[3] Hossam Faris, Ibrahim Aljarah, and Seyedali Mirjalili. Improved monarch butterfly optimization for unconstrained global search and neural network training. *Applied Intelligence*, 48(2):445–464, 2018.

[4] Mohammad Al Jabery and Marshall Zumberg. General and special education systems in jordan: Present and future perspectives. *International Journal of Special Education*, 23(1):115–122, 2008.

[5] Ministry of Higher Education ministry of higher education, 2018.

[6] Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Preprocessing and analyzing educational data set using x-api for improving student's performance. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5. IEEE, 2015.

[7] Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8):119–136, 2016.

[8] Ryan SJD Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM| Journal of Educational Data Mining*, 1(1):3–17, 2009.

[9] William W Cohen. Fast effective rule induction. In *Machine Learning Proceedings 1995*, pages 115–123. Elsevier, 1995.

[10] Eibe Frank and Ian H Witten. Generating accurate rule sets without global optimization. *University of Waikato Researc*, 1998.

[11] Ryan J Urbanowicz and Jason H Moore. Learning classifier systems: a complete introduction, review, and roadmap. *Journal of Artificial Evolution and Applications*, 2009:1, 2009.

[12] Eibe Frank and A Mark. Hall, and ian h. witten, 2016. *The WEKA Workbench. Online Appendix for: Data Mining: Practical Machine Learning Tools and Techniques*, 2016.

[13] Sai Sumathi and SN Sivanandam. *Introduction to data mining and its applications*, volume 29. Springer, 2006.

[14] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[15] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.*, 33(1):135–146, July 2007.

[16] Kenneth R Koedinger, Sidney D'Mello, Elizabeth A McLaughlin, Zachary A Pardos, and Carolyn P Rose. Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(4):333–353, 2015.

[17] Abdulmohsen Algarni. Data mining in education. *International Journal of Advanced Computer Science and Applications*, 7(6):456–461, 2016.

[18] Dorina Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

[19] Amirah Mohamed Shahiri, Wahidah Husain, et al. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422, 2015.

[20] Scott S Meit, Nicole J Borges, Barbara A Cubic, and Hugo R Seibel. Personality differences in incoming male and female medical students. *Online Submission*, 2004.

[21] Ali Simsek and Jale Balaban. Learning strategies of successful and unsuccessful university students. *Online Submission*, 1(1):36–45, 2010.

[22] Ardita Permata Alfiani, Febriana Ayu Wulandari, et al. Mapping student's performance based on data mining approach (a case study). *Agriculture and Agricultural Science Procedia*, 3:173–177, 2015.

[23] Mr MN Quadri and NV Kalyankar. Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 2010.

[24] D Magdalene Delighta Angeline et al. Association rule generation for student performance analysis using apriori algorithm. *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, 1(1):12–16, 2013.

[25] M Mayilvaganan and D Kalpanadevi. Comparison of classification techniques for predicting the performance of students academic environment. In *Communication and Network Technologies (ICCNT), 2014 International Conference on*, pages 113–118. IEEE, 2014.

[26] Esther Y Strahan. The effects of social anxiety and social skills on academic performance. *Personality and individual differences*, 34(2):347–366, 2003.

[27] Mohd Shahizan Othman, Mahdi Alhaji Musa, et al. The improvement of students' academic performance by using social media through collaborative learning in malaysian higher education. *Asian Social Science*, 10(8):210, 2014.

[28] Chiara Masci, Geraint Johnes, and Tommaso Agasisti. Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, 269(3):1072–1085, 2018.

[29] Drew O'Brien. Does financial aid affect academic performance? the effects of student aid on gpa. *Van Wylen Library*, 2015.

[30] Gloria Crisp, Amanda Taggart, and Amaury Nora. Undergraduate latina/o students: A systematic review of research identifying factors contributing to academic success outcomes. *Review of Educational Research*, 85(2):249–274, 2015.

[31] Joanne Gikas and Michael M Grant. Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media. *The Internet and Higher Education*, 19:18–26, 2013.

[32] Paul A Tess. The role of social media in higher education classes (real and virtual)–a literature review. *Computers in Human Behavior*, 29(5):A60–A68, 2013.

[33] Ritesh Chugh and Umar Ruhi. Social media in higher education: A literature review of facebook. *Education and Information Technologies*, 23(2):605–616, 2018.

[34] Sedigheh Moghavvemi, Ainin Sulaiman, Noor Ismawati Jaafar, and Nafisa Kasem. Social media as a complementary learning tool for teaching and learning: The case of youtube. *The International Journal of Management Education*, 16(1):37–42, 2018.

[35] Vladlena Benson and Stephanie Morgan. Measuring the social impact: how social media affects higher education institutions. In *Social Media Marketing: Breakthroughs in Research and Practice*, pages 1167–1184. IGI Global, 2018.

[36] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[37] Surveymonkey.

[38] SB Kotsiantis, Dimitris Kanellopoulos, and PE Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117, 2006.

[39] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000.

[40] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[41] Ali Asghar Heidari, Ibrahim Aljarah, Hossam Faris, Huiling Chen, Jie Luo, and Seyedali Mirjalili. An enhanced associative learning-based exploratory whale optimizer for global optimization. *Neural Computing and Applications*, pages 1–27, 2019.

[42] Hossam Faris, Ibrahim Aljarah, Seyedali Mirjalili, Pedro A Castillo, and Juan Julián Merelo Guervós. Evolopy: An open-source nature-inspired optimization framework in python. In *IJCCI (ECTA)*, pages 171–177, 2016.

[43] Sarah Shukri, Hossam Faris, Ibrahim Aljarah, Seyedali Mirjalili, and Ajith Abraham. Evolutionary static and dynamic clustering algorithms based on multi-verse optimizer. *Engineering Applications of Artificial Intelligence*, 72:54–66, 2018.