Mining Student Information System Records to Predict Students' Academic Performance

تعدين سجلات نظام معلومات الطلبة للتنبؤ بأداءهم الأكاديمي

by

AMJED TARIQ MOHAMMAD ABU SAA

A dissertation submitted in fulfilment
of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)
at

The British University in Dubai

November 2018

Declaration

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright materials falls within the limits permitted by international copyright conversations.

I understand that a copy of my research will be deposited in the university library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purpose of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the university to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

_____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

An increasing interest has arisen during the past decade to identify the most important factors influencing students' performance in higher education, especially by using predictive data mining techniques. This field of research is usually identified as educational data mining. Educational Data Mining (EDM) is the field of study that is concerned about mining useful patterns and predicting student's academic performance in the field of education. Researchers in this field tend to study different types of students' factors and attributes that affect their performance and learning outcomes. In this dissertation, 36 research articles out of a total of 420 from 2009 to 2018 were critically reviewed and analyzed by applying a systematic literature review approach. As well as, this dissertation provides a predictive data mining study targeted towards the use of student information systems' data to predict students' academic performance. A gap between student information systems and data mining was identified and addressed in this study, which suggests connecting both worlds together creating an intelligent system that is capable to predict student's failures and low academic performance according to relevant students' attributes. The main aim of this study is to identify the most commonly studied factors that affect the students' performance, as well as, the most common data mining techniques applied to identify these factors. Accordingly, this dissertation generated a dataset from a student information system from a local university in the United Arab Emirates. The dataset included 34 attributes of student's related information, and was having a data size of more than 56,000 records. Empirical results showed that four types of student attributes are responsible for academic performance prediction, including, students' demographics, students' previous performance information, course and instructor information as well as some student general information. Additionally, the results also indicated that the most common data mining techniques used to predict and classify students' factors are decision trees, Naïve Bayes, and artificial neural networks. Finally, a set of data-mining models was compared in order to identify the most suitable one for predicting students' academic performance from student information systems.

Keywords: Educational Data Mining; students' factors; students' academic performance; systematic review; data mining techniques; student information systems.

ملخـــــــــص

نشأ اهتمام متزايد خلال العقد الماضي لتحديد أهم العوامل التي تؤثر على أداء الطلاب في التعليم العالي، وخاصة باستخدام تقنيات التعدين عن البيانات التنبؤية. عادة ما يتم وصف هذا المجال من البحوث باسم تعدين البيانات التربوية. تعدين البيانات التربوية (EDM) هو مجال الدراسة الذي يهتم بأنماط التعدين المفيدة والتنبؤ بالأداء الأكاديمي للطالب في مجال التعليم. يميل الباحثون في هذا المجال إلى دراسة أنواع مختلفة من عوامل الطلاب والسمات التي تؤثر على أدائهم ونتائج التعلم المرجوة. في هذه الرسالة، تمت مراجعة وتحليل 36 مقالة بحثية من إجمالي 420 من العام 2009 إلى 2018 وذلك من خلال تطبيق منهجية مراجعة الأدبيات المنهجية. بالإضافة إلى ذلك، توفر هذه الرسالة دراسة تنبؤية في البيانات التنبؤية تستهدف استخدام بيانات أنظمة معلومات الطلاب للتنبؤ بالأداء الأكاديمي للطلاب. وقد تم التعرف على فجوة بين أنظمة معلومات الطلاب وتعدين البيانات التي تم تناولها في هذه الدراسة، والتي تقترح ربط العالمين معًا بإنشاء نظام ذكي قادر على توقع حالات فشل الطلاب وانخفاض مستوى الأداء الأكاديمي وفقًا لسمات الطلاب ذات الصلة. يتمثل الهدف الرئيسي من هذه الدراسة في تحديد العوامل الأكثر شيوعًا والتي تؤثر على أداء الطلاب، بالإضافة إلى تقنيات التعدين الأكثر شيوعًا المستخدمة لتحديد هذه العوامل. بناءً على ذلك، أنتجت هذه الرسالة مجموعة بيانات من نظام معلومات الطلبة من جامعة محلية في الإمارات العربية المتحدة. تضمنت مجموعة البيانات 34 سمة من معلومات الطالب ذات الصلة، وكان حجم بياناتها أكثر من 56000 سجل. أظهرت النتائج التجريبية أن أربعة أنواع من سمات الطالب مسؤولة عن التنبؤ بالأداء الأكاديمي، بما في ذلك الخصائص الديموغرافية للطلاب ومعلومات الأداء السابقة للطلاب ومعلومات المواد الدراسية والمعلم أو المحاضر بالإضافة إلى بعض المعلومات العامة للطلاب. بالإضافة إلى ذلك، أشارت النتائج أيضًا إلى أن تقنيات التعدين الأكثر شيوعًا المستخدمة للتنبؤ بعوامل الطلاب وتصنيفها هي أشجار القرار، Naïve Bayes، والشبكات العصبية الاصطناعية. وأخيرًا ، تمت مقارنة مجموعة من نماذج استخراج البيانات من أجل تحديد النموذج الأفضل للتنبؤ بالأداء الأكاديمي من أنظمة معلومات الطلاب.

Dedication

With all pride and pleasure, I dedicate my thesis to everyone who wished me success and excellence from the depth of their hearts, and to everyone helped and supported me, And to all who prayed for my success and my happiness.

I dedicate my work to my family, my kids and my beloved wife, who has always supported me morally, physically and emotionally, you are always my real success and I wish us greater success than mine, and I hope to be a source of pride and lasting love for you.

I’m blessed with everything I need. I am working hard towards everything I want, and most of all I appreciate and thank God for what I have.

Acknowledgement

I will always remember to thank God for what he has given me, for blessing me and giving me the patience to finish my thesis that I hope it could be something useful for the world.

I would like to express my sincere gratitude to my supervisor Prof. Khaled Shaalan for the continuous support of my MSc study and related research, for his patience, motivation, and immense knowledge. His guidance helped during the time of research and writing of this dissertation. I could not have imagined having a better supervisor and mentor for my MSc study.

Table of Contents

List of Tables

List of Figures

1.

Chapter One: Introduction

1.1. Educational Data Mining

An increasing interest has arisen during the past decade to identify the most important factors influencing students' performance in higher education, especially by using data mining methods and techniques. This field of research are usually identified as educational data mining (EDM) (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018). The reason behind this interest are the applicability of such research in helping to identify low performing students early enough to overcome their difficulties in learning and improve their learning outcomes accordingly, which serves the institutional goals of providing high quality education ecosystems. In addition, EDM is fast becoming an important field of research due to its ability to extract a new knowledge from a huge amount of students' data (Wook, Yusof, & Nazri, 2017). This dissertation is equally interested in this topic, and our objective is to explore and review papers from the past decade that are in the context of educational data mining and identifies the main factors influencing students' performance in higher education. As well as, we aim to provide an Educational Data Mining implementation on an extracted data from a student information system. EDM is defined by the Educational Data Mining community website (www.educationaldatamining.org) as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in".

Researchers in this field tend to study different types of students' factors and attributes that affect students' performance and learning outcomes. (Shahiri, Husain, & Rashid, 2015) conducted a systematic literature review on predicting students' performance using

data mining techniques. The review tackled many subjects, one of which was to identify the important attributes used in predicting the students' performance. Results showed that cumulative grade point average and internal assessments are the most frequent attributes used for predicting the students' performance. Furthermore, other important attributes were also identified, including students' demographic and external assessments, extra-curricular activities, high school background, and social interaction network. Additionally, the results showed that Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine were the most frequently used data mining techniques in the EDM research. Besides, (Peña-Ayala, 2014) conducted a survey and meta-analysis of recent studies related to EDM. The results indicated that 60% of EDM research articles have used predictive data mining approaches as opposed to 40% which have used the descriptive approaches. Furthermore, the results also showed that classification and clustering were the most typical techniques used by EDM research. Additionally, Bayes theorem, decision trees, instances-based learning (IBL), and hidden Markov model (HMM) were found to be the most popular methods used by EDM research. Furthermore, (C. Romero & Ventura, 2007) carried out a review study aiming to analyze the application of data mining for different educational systems: traditional system, web-based courses, content management systems, and intelligent web-based systems. The results suggested investigating the applicability of using data mining techniques for e-learning systems.

1.2. Data Mining and Student Information Systems

In the past two decades, we witnessed a vast technological advancement in the area of computers and information systems (S.A. Salloum, Al-Emran, Monem, & Shaalan, 2018). Educational institutions took the advantage of this advancement and employed it to digitalize most of its educational and academic information, and transactions (Yukselturk et al., 2014). Currently, we hardly find any higher education institution

without a student information system, where most of the students' information resides (Cristbal Romero & Ventura, 2010). Student Information Systems store a huge amount of data about students, such as student's demographics, courses' information, instructors' information, students' class attendance, students' grades, and many more. There is a high potential in those systems that enables them to be used in educational data mining to predict students' performance based on their data in the student information system (Kiron et al., 2012). This study addresses this matter, where we try to collect as many students' information as we can from an information system of a university in the United Arab Emirates (UAE), and we try to analyze and predict the student performance based on the real data. In addition, we try to identify the most appropriate data mining technique for use in prediction.

Analyzing and predicting students' academic performance has been an emerging topic in the past two decades, and has been an interest for many researchers and scholars (Cristbal Romero & Ventura, 2010; Xing, Guo, Petakovic, & Goggins, 2015; Yukselturk et al., 2014). With no exception, this study focuses on this subject. The most important feature of such study is to learn from past data, and help instructors and university management understand the factors affecting their students' performance, so they will be able to focus on low performing students and help them overcome their weaknesses and improve quality (Bienkowski, Feng, & Means, 2012; Xing et al., 2015).

A variety of data mining techniques has been employed on the collected data from the student information system to build predictive models for students' academic performance prediction. Based on the findings of a recent systematic literature review on educational data mining (EDM) (Papamitsiou & Economides, 2014; Shahiri et al., 2015), results indicate that the most frequently used data mining techniques by the EDM community are Decision Trees (DT), Naive Bayes (NB), Artificial Neural Networks

(ANN), Support Vector Machine (SVM), and Logistic Regression. Henceforth, we will be focusing on these set of techniques when selecting our data mining (DM) algorithms.

A study conducted by (Márquez-Vera et al., 2016) studied 17 attributes of students, in which, nine of them were extracted from a student information system of a high school. The studied attributes are related to the student scores in specific subjects, students' attendance, and other attributes. It was found that the GPA in secondary school, classroom/group enrolled, and the number of students in the group/class, among other attributes were the most significant and had a large impact on the student performance in a high school.

Another study carried out by (Asif, Merceron, Ali, & Haider, 2017) conducted an educational data mining research on admission data trying to build an early warning system by predicting low performing students as early as possible, and provide them with the possible opportunities to improve their performance. The study had a simple dataset, consisting of eight attributes only. All the attributes were related to the students' high school marks as well as the first and second year of university. All data were collected from the university's student information system where the study took place. The study used multiple classification techniques to model their system, including Decision Trees, Rule Induction, K-NN, Naïve Bayes, etc. Results showed that all the attributes used in data mining had a significant impact on students' academic performance, suggesting that all previous grades and scores achieved by a student is always a successful predictor of a student future grade.

Fernandes and his colleagues (Fernandes et al., 2018) conducted a predictive analysis of academic performance of public school students. The authors studied the effect of multiple students' attributes collected from two different datasets acquired at the beginning of the semester, and after two months from starting the semester. The second

dataset included the same attributes of the first dataset, and added multiple attributes that became available after two months of studying, such as grades, absences, and subjects' information. Results showed that after adding these attributes, the importance of attributes from the first dataset became less significant. However, not all attributes of the second dataset bypassed the importance of first dataset's attributes. Finally, significant attributes according to their scale of importance was: 1) grades, 2) student neighborhood, 3) school, 4) school subjects, 5) absences, 6) student city, and 7) age.

A study by (Huang & Fang, 2013) conducted a comparative research of predictive models of students' academic performance. The study was very specific, in which it attempted to predict the students' performance on a specific course (Engineering Dynamics). The researcher studied a limited number of attributes extracted from the student information system. Specifically, attributes included students' last GPA, grades on related subjects, and scores of mid-term exams of the engineering dynamics course. Results showed that there were insignificant differences between the four DM models applied on the data when applied on all students. However, when applied to one student only, SVM was the most appropriate predictive model to be used. Additionally, it was found that the GPA is the only effective predictor variable among the other attributes used in the modeling when it is applied to all students; however, when predicting individual student's performance, all attributes make a difference.

1.3. Research Motivation & Objective

It is observed that there is a lack of agreement among most of the research articles regarding the factors that are believed to influence the students' performance. Thus, there is a clear need to identify the most important and most studied factors that were found significant and truly affect the students' performance from the increasing amount of the literature available in the EDM field.

The aim of this dissertation is to investigate the literature related to EDM and to identify the most important and most studied factors influencing the students' performance in higher education, as well as, to generate a generalized set of factors and attributes that are believed to affect the students' performance and learning outcomes in the higher education sector. In order to perform this study, we will review the literature using systematic literature review (SLR) method.

The second main goal of the present study is to build a predictive model based on student information system's data and select the most appropriate data mining technique and algorithm to predict students' academic performance. This research uses a list of 34 attributes of data extracted from a student information system of a local university in the UAE. It is worth mentioning that the selected university has a challenging environment for research. First, it is segregated by gender, which means male and female students are not mixed up in classes, and even on campus, for culture and conservative reasons. Second, it has a variety of students coming from more than 100 nations. In fact, the UAE itself has an interesting environment of research, because of its unique characteristics and demographics. The country has vastly developed over the past 40 years, as well as, the national population are considered a minority (less than 20% of total population), and Islamic culture and modesty are dominant in the country, especially in schooling. Identically, the university demographics include a mixture of both local and other nationalities, mostly Arabs from the Middle East region who were raised in the country with their working parents, and few international students. Therefore, our research falls in such a challenging environment with many different students' backgrounds and many diversities.

1.4. Research Questions

This dissertation attempts to address the following research questions:

Research Question 1: What are the most common and most frequently used factors affecting students' performance in higher education?

Research Question 2: What are the most common and most frequently used data mining techniques used to analyze and predict students' academic performance?

Research Question 3: What is the most appropriate data mining technique/algorithm that has the best results for predicting students' academic performance using real data extracted from a student information system?

Research Question 4: What are the main predictors of students' academic performance among the attributes selected from a student information system?

1.5. Research Methodology

This dissertation is split into main two parts, the first part are planned to provide a systematic literature review of prior educational data mining research studies, and answer the first two research questions defined in the previous section (1.5.). The SLR study aims to review state of the art research papers in the EDM domain. The SLR study, provided in chapter two, follows a standardized SLR methodology inspired by the work conducted by (Kitchenham et al., 2009). The methodology followed in the SLR study is split into three main phases as described in Figure 1.

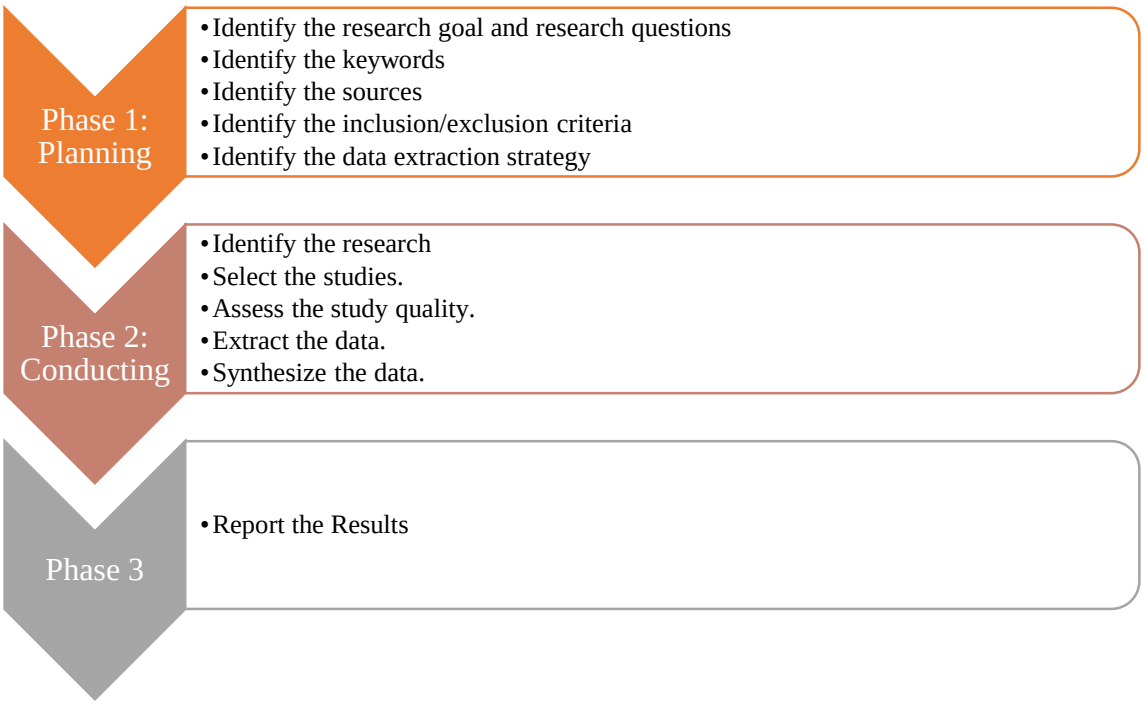

Figure 1: Phases of SLR Study

Thereupon, the second part of the dissertation is planned to use the findings of the conducted SLR, and employ it in a data mining implementation. The dataset used in this part of the dissertation is extracted from a student information system of a local university in the United Arab Emirates. The dataset contains students' attributes that is compliant with the findings of the SLR study on the most common students' factors affecting academic performance among the EDM community. Additionally, the data mining techniques employed in this part follows the findings of the SLR on the most common used data mining techniques by the EDM community as well. Finally, results and findings of both parts are summarized in chapter four and concluded in chapter five.

1.6. Dissertation Structure

This dissertation is divided into five chapters. Chapter two provides a systematic literature review which reviews and identifies the findings of prior research studies in order to identify the set of most common factors affecting students' academic performance in the literature, as well as, the most frequently used data mining techniques used in educational

data mining research. The findings of the systematic literature review will be used later in Chapter three to apply the most frequently used data mining techniques on a set of attributes that are also bound with the findings of the SLR's most common factors affecting students' academic performance. Chapter four provides answers to the four research questions identified in section 1.5. Finally, Chapter five provides the conclusion and future prospects of this dissertation research study.

2. Chapter Two: Systematic Literature Review

2.1. Methodology

In this chapter, we employed a standard SLR methodology, which is inspired by the work conducted by (Kitchenham et al., 2009).

Systematic Literature Review is one of the most common approaches used for literature review, and it serves our objective in this research which is supposed to provide a summary of studies related to EDM and identify the factors affecting students' academic performance in higher education that are most commonly used and were found to have significant effects and results. SLR provides an appropriate procedure and a framework that improves the quality of research papers, literature reviews, and evaluations (Budgen & Brereton, 2006). Using a clear SLR protocols effectively guides researchers throughout the process of the review, as well as, improves the methodological transparency of the review and enables future replication (Mallett, Hagen-Zanker, Slater, & Duvendack, 2012). SLR has many advantages over simple and unstructured literature review methods, as it is more likely to be considered reliable and unbiased, as well as, information gathered in SLRs are with a larger scale and wider sources (Kitchenham & Charters, 2007). There are three main phases of SLR: planning, conducting, and reporting. Following are the sub-steps in each phase as outlined by (Kitchenham et al., 2009) and (Al-Araibi, Mahrin, & Yusoff, 2016) that should be performed in an SLR study:

Phase 1: Planning

1. Identify the research goal and research questions.
2. Identify the keywords.
3. Identify the sources.
4. Identify the inclusion/exclusion criteria.

5. Identify the data extraction strategy.

Phase 2: Conducting

1. Identify the research.
2. Select the studies.
3. Assess the study quality.
4. Extract the data.
5. Synthesize the data.

Phase 3: Reporting the Results

2.2. Phase 1: Planning

In this section, we describe in detail the steps of phase 1 (i.e., planning) that were conducted in our study, which includes, identify the research goal and research questions, identify the keywords, identify the sources, identify the inclusion/exclusion criteria, and identify the data extraction strategy.

2.2.1. Identify the research goal and research questions

Our objective in this chapter is to systematically review relevant literature through a Systematic Literature Review (SLR) process (Al-Araibi et al., 2016; Kitchenham & Charters, 2007; Kitchenham et al., 2009), and our research questions are provided as follows:

- What are the most common and most frequently used factors affecting students' performance in higher education?
- What are the most common and most frequently used data mining techniques used to analyze and predict students' academic performance?

2.2.2. Identify the keywords

Our search keywords were mostly driven by the research questions stated in the previous subsection. After identifying the search keywords, we had to prepare a search string that should work with the search engines of the libraries to be searched which will be identified in the next section. The following search string were prepared:

("data mining" OR "educational data mining") AND ("factors affecting student performance" OR "predicting students performance")

As it can be seen in the search string, the term "predicting students' performance" was added to the search string even though it did not appear in the research questions; this is because we have noticed in the planning stage that there are a lot of research that included this term in their titles and/or abstracts which identifies that the research are related to EDM, and they predict the students' performances based on other attributes and factors which they collect in their studies, in which, this clearly satisfies our objective in this step.

2.2.3. Identify the sources

The following online library databases and search engines were selected to be searched for our SLR: Science Direct, EBSCO, ProQuest, JSTOR, Taylor & Francis Online.

2.2.4. Identify the inclusion/exclusion criteria

Our inclusion/exclusion criteria are shown in Table 1. Each study found in the search results must meet these criteria in order to be included in our SLR.

Inclusion Criteria	Exclusion Criteria
a. Must meet the research keywords conditions.	a. Doesn't meet the research keywords conditions.
b. Must be classified as a data mining or machine learning research.	b. Not classified as a data mining or machine learning research.
c. Must include the studied factors.	c. Does not include the studied factors.
d. Full text paper must be available and accessible, and must not be accessible via arXiv.	d. Full text paper is not available nor accessible, nor accessible via an arXiv.

e. Must not be a review paper.	e. Is a review paper.
f. Must be published in the last decade (i.e., between 2009 and 2018)	f. Published but its date exceeds 10 years ago (i.e., earlier than 2009)
g. Must be written in English language.	g. Not written in English language.

Table 1: Inclusion/exclusion criteria

2.2.5. Identify the data extraction strategy

In this study, the data were collected based on the fields described in Table 2.

Item	Item Description
Paper ID	An ID number is assigned to each research paper in order to be easily referenced during the review.
Source	The database source of the research paper.
Paper Title	The title of the research paper.
Journal	The journal that published the research paper.
Author	The author of the research paper.
Year	The paper publication year.
Country of Study	The country in which the study of the research paper was undertaken.
Studied Factors	The list of factors that were studied in the research paper.
Factors Category(s)	The categories of the factors in the previous field, such as: students demographics, students social information, e-Learning activities, etc.
Factors Found Significant	The list of factors that were found significant by the researchers of this study out of the full list of studied factors.
Data Mining Approach(es)	The data mining technique used in the research paper, such as: classification, clustering, etc.
Data Mining Algorithms	The data mining algorithms that were used in the research paper, such as, decision trees, SVM, K-Means clustering, etc.
Data Collection Technique(s)	Technique(s) that were used to collect data in the research paper, such as, surveys, student information systems data, e-Learning system data, etc.
Data Set Size	The size of the data set that were used in the research paper.

Table 2: Data Layout

Research papers that did not have one or more fields from the data described in Table 2, were also excluded from the study. In other words, research papers that exhibited full details were covered and included in the study.

2.3. Phase 2: Conducting the Review

In this section, we describe in detail the steps of phase 2 (i.e., Conducting) that were conducted in our study, which includes, identify the research, select the studies, assess the study quality, extract the data, synthesize the data.

2.3.1. Identify the research

In this step, we started searching the online libraries' databases with the aforementioned search string. The initial search results returned by the search engines are illustrated in Table 3.

Online Library Database / Search Engine	Results
Science Direct	48
EBSCO	80
ProQuest	34
JSTOR	201
Taylor & Francis Online	57
Total	420

Table 3: Initial Search Results

2.3.2. Select the studies

In this step, we select the research papers that meets our inclusion/exclusion criteria, as well as, further investigate the selected papers' contents to verify their eligibility for selection. We applied automatic and semi-automatic paper selection. As a result of the automatic selection, we found 218 duplicate results in our search, 57 did not meet the research keywords and were not related to the subject of this SLR study, 19 were not a data mining research, 23 were not free access nor available, or an arXiv paper, 21 were

review papers, and 34 non-English papers. Furthermore, research dates were set as a search filters in the search engine and were excluded from the initial search results. Finally, we were left with 48 research papers. Then, we applied the semi-automatic paper selection. Consequently, after reading the full text research papers and extracting the SLR data, we found 12 papers that did not include the studied factors that affect the students' performance; hence, they were also excluded from the list. Therefore, the final data set size of our SLR is 36 research papers. Table 4 lists all the 36 research papers selected for this SLR.

Paper ID	Source	Journal	Author
RP1	Science Direct	Computers in Human Behavior	(Xing et al., 2015)
RP2	Science Direct	Computers & Education	(Asif et al., 2017)
RP5	Science Direct	Journal of Business Research	(Fernandes et al., 2018)
RP7	Science Direct	Computers & Electrical Engineering	(Burgos et al., 2018)
RP9	Science Direct	Computers & Education	(Lara, Lizcano, Martínez, Pazos, & Riera, 2014)
RP10	EBSCO	Expert Systems	(Gómez-Rey, Fernández-Navarro, & Barberà, 2016)
RP12	ProQuest	Informatics in Education	(Jiang, Javaad, & Golab, 2016)
RP16	ProQuest	Applied Intelligence	(Márquez-Vera, Cano, Romero, & Ventura, 2013)
RP22	EBSCO	Expert Systems	(Márquez-Vera et al., 2016)
RP23	JSTOR	Journal of Educational Technology & Society	(Abdous, He, & Yen, 2012)
RP25	JSTOR	Journal of Educational Technology & Society	(Hung, Hsu, & Rice, 2012)

RP28	Science Direct	Knowledge-Based Systems	(S. Kotsiantis, Patriarcheas, & Xenos, 2010)
RP33	Science Direct	Applied Soft Computing	(Zafra & Ventura, 2012)
RP34	Science Direct	Computers & Education	(Cristóbal Romero, López, Luna, & Ventura, 2013)
RP35	EBSCO	Applied Stochastic Models in Business & Industry	(Costantini, Linting, & Porzio, 2010)
RP36	EBSCO	Expert Systems	(Gamulin, Gamulin, & Kermek, 2016)
RP43	ProQuest	The Artificial Intelligence Review	(S. B. Kotsiantis, 2012)
RP44	Science Direct	Computers in Human Behavior	(Hu, Lo, & Shih, 2014)
RP67	JSTOR	European Journal of Open, Distance and E-Learning	(Yukselturk et al., 2014)
RP69	JSTOR	Journal of Computer Science	(Abazeed & Khder, 2017)
RP75	Google Scholar	International Journal of Advanced Computer Science and Applications	(Abu Saa, 2016)
RP81	JSTOR	Indian Journal of Science and Technology	(C. Anuradha Bharathiar, 2015)
RP94	JSTOR	The Electronic Journal of Information Systems in Developing Countries	(Mwalumbwe & Mtebe, 2017)
RP113	Science Direct	Computers & Education	(Macfadyen & Dawson, 2010)
RP120	Google Scholar	International Journal of Advanced Computer Science and Applications	(Baradwaj & Pal, 2012)
RP123	Google Scholar	World Journal of Computer Application and Technology	(Badr El Din Ahmed & Sayed Elaraby, 2014)

RP127	Google Scholar	International Journal of Computer Science and Information Technologies	(Pandey & Pal, 2011)
RP128	Google Scholar	International Journal of Computer Science and Information Security	(Bhardwaj & Pal, 2012)
RP130	Google Scholar	International Journal of Innovative Technology & Creative Engineering	(S. Yadav, Bharadwaj, & Pal, 2012)
RP136	Google Scholar	World of Computer Science and Information Technology Journal	(S. K. Yadav & Pal, 2012)
RP142	Science Direct	Computers & Education	(Araque, Roldán, & Salguero, 2009)
RP174	ProQuest	International Conference on Digital Information and Communication Technology and its Applications	(Zhou, Zheng, & Mou, 2015)
RP198	Science Direct	Computers & Education	(Cerezo, Sánchez-Santillán, Paule-Ruiz, & Núñez, 2016)
RP220	ProQuest	The International Journal of Information and Learning Technology	(Chamizo-Gonzalez, Cano-Montero, Urquia-Grande, & Muñoz-Colomina, 2015)
RP277	EBSCO	Journal of AI and Data Mining	(Hasheminejad & Sarvmili, 2018)
RP323	Science Direct	Computers & Education	(Huang & Fang, 2013)

Table 4: Selected Research Papers

2.3.3. Assess the study quality

In order to assess the quality of the selected papers in the previous subsection, we have to answer the questions in Table 5for each paper in the data set.

#	Question
Q1	Are the study aims clearly stated?
Q2	Is the research described adequately?
Q3	Does the study explore diversity of perspectives and contexts?
Q4	Do the objectives lead to conclusions clearly?
Q5	Are the findings important?
Q6	Are negative findings presented?
Q7	Do the researchers explain the consequences of any problems?
Q8	Does the study add to your knowledge or understanding?
Q9	Do the results add to the literature?

Table 5: Paper Quality Assessment Questions (Kitchenham & Charters, 2007)

Correspondingly, the answers of the questions accept three scores: Yes (1), Partially (0.5), and No (0). Summing up the scores of all questions for each study will result in a cumulative quality score for each paper out of 9. The result is then converted to a percentage, e.g. 7 out of 9 is 77.78%. Table 6 shows the cumulative quality scores for all the papers in our dataset and the percentage results.

Paper ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Quality Score	Percentage
RP1	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP2	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP5	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP7	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP9	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP10	1	1	1	1	1	0	1	1	1	8	88.89%
RP12	1	1	1	1	1	0	0	1	1	7	77.78%
RP16	1	1	1	1	0.5	0	0	0.5	1	6	66.67%
RP22	1	1	1	1	1	0	1	1	1	8	88.89%
RP23	1	1	1	1	1	0	0.5	1	0.5	7	77.78%
RP25	1	1	1	1	0.5	0	0.5	0.5	0.5	6	66.67%

RP28	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP33	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP34	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP35	1	1	1	1	0.5	0	0	0.5	0.5	5.5	61.11%
RP36	1	1	1	1	0.5	0	0.5	0	0.5	5.5	61.11%
RP43	1	1	1	1	0.5	0	0	0.5	0.5	5.5	61.11%
RP44	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP67	1	1	1	0.5	0.5	0	0.5	0	1	5.5	61.11%
RP69	1	1	1	0.5	1	0	0	0	1	5.5	61.11%
RP75	1	1	1	1	1	0	0	1	1	7	77.78%
RP81	1	0	0.5	1	0.5	0	0.5	0.5	0.5	4.5	50.00%
RP94	1	0.5	0.5	1	0.5	0	1	0	1	5.5	61.11%
RP113	1	1	1	1	1	0	0.5	1	1	7.5	83.33%
RP120	1	0.5	0.5	0.5	0	0	1	1	0.5	5	55.56%
RP123	1	0.5	0.5	1	1	0	0	1	1	6	66.67%
RP127	1	0.5	0.5	1	1	0	0	0.5	1	5.5	61.11%
RP128	1	0.5	0.5	0.5	1	0	0	0.5	0.5	4.5	50.00%
RP130	1	0.5	0.5	0.5	0.5	0	0	0.5	0.5	4	44.44%
RP136	1	0.5	0.5	0.5	0.5	0	0	0.5	0.5	4	44.44%
RP142	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP174	1	1	1	1	1	0	0	1	1	7	77.78%
RP198	1	1	1	1	1	0	0.5	0.5	1	7	77.78%
RP220	1	1	1	1	0.5	0	0.5	0.5	1	6.5	72.22%
RP277	1	1	1	1	0.5	0	0	0.5	0.5	5.5	61.11%
RP323	1	1	1	1	1	0	0.5	0.5	1	7	77.78%

Table 6: Quality Scores and Percentages

As shown in Table 6, 34 out of 36 papers achieved a score of 4.5 (50%) or higher, and 2 papers were scored 4/9 (44.44%) due to their low quality and poor content. The highest scored papers are RP10 and RP22 with a score of 8/9 (88.9%). Consequently, the two

low-scored papers (RP130, RP136) were removed from the SLR process, and the remaining 34 were kept for the subsequent steps.

2.3.4. Extract the data

In this step, the required data for our SLR study will be extracted from the selected papers according to the data layout in Table 2. We focused in this step on finding the factors that affect the students' performance, and most importantly, those that were found significant by the researchers in their papers, as well as, the data mining techniques and algorithms used by the researchers in their data mining research. These data will help us to get useful insights and results that will empower us to answer our research questions. Table 7 summarizes the extracted data for each paper in our dataset.

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP1	Students e-Learning activity	• Chat logs of all messages that students send to each other in the group. • Awareness records of actions of erasing the chat messages on the chat bar. • Geogebra logs of information on how students virtually construct a geometry artifact (adding a point or updating a segment). • System logs of students joining a virtual room, leaves a virtual room or views different tabs. • WhiteBoard logs of more specific actions on how tools are being used in the white board areas such as resizing objects or creating a textbox.	Classification	e-Learning System Logs
RP2	Students Previous	• High School Marks (total and subject specific) • first and second year university courses' marks	1.Classification 2.Clustering	Admission Data

	Grades & Class Performance			
RP5	1. Students Environment 2. Students Demographics 3. Students Previous Grades & Class Performance	• Grades for the first two months • Student's place of residence – neighborhood • School name • School subjects • Absences • Student's place of residence – city • Age	Classification	Student Information System Data
RP7	Students e-Learning activity	• 12 Assessment activities from e-Learning system • Teaching Schedule	Classification	e-Learning System Logs
RP9	Students e-Learning activity	• Number of virtual classroom accesses by the student in the week in question • Number of different days of the week on which the student accesses the virtual classroom • Whether or not the resource has been visualized in the week in question • Number of times that the student has visualized the resource in the week in question	Classification	e-Learning System Logs
RP10	1. Instructor Attributes 2. Students Previous Grades & Class Performance 3. Course Attributes	• Instructor's knowledge • Instructor's effective use of the class hours • Instructor's coherence with lesson plan • Openness and respect of the instructor to students' views • Instructor's positive approach to students • Instructor readiness for classes • Instructor explanations about the course and instructor helpfulness	Classification	Course Evaluations Surveys

RP12	1. Instructor Attributes 2. Course Attributes	• Instructor's organization and clarity • Instructor's response to questions • Instructor's visual presentation • Instructor's encouragement to think independently • Instructor's attitude towards teaching • Professor-class relationship • Difficulty of concepts covered • Contribution of assignments to understanding of concepts • How well tests reflect the course material • Attendance (the number of evaluations received divided by course enrolment)	Classification	Course Evaluations Surveys
RP16	1. Students Previous Grades & Class Performance 2. Students Demographics 3. Students Social Data	• Scores in specific subjects • Level of motivation • GPA in secondary school • Age • Number of brothers/sisters • Classroom/group • Smoking habits • Studying in group • Marital status • Time spent doing exercises	Classification	1. Surveys 2. Student Information System Data
RP22	1. Students Previous Grades & Class Performance 2. Students Demographics	• GPA in secondary school • Classroom/group enrolled • Number of students in the group/class • Age • Attendance during morning/evening sessions • Having a job • Mother's level of education	Classification	1. Surveys 2. Student Information System Data

	3. Students Social Data			
RP23	Students e-Learning activity	• Students activity data from an online video e-learning system: • Number of questions • Number of chat messages • Total login times • Final grade	1.Classification 2.Clustering	e-Learning System Logs
RP25	1. Students e-Learning activity 2. Students Demographics 3. Course Evaluations	1. Students e-Learning activity: • Average frequency of logins per course • Average frequency of tab accessed per course • Average frequency of module accessed per course • Average frequency of clicks per course • Average frequency of course accessed per course • Average frequency of page accessed per course • Average frequency of course content accessed per course • Average number of discussion board entries per course 2. Students Demographics: Age, gender, graduation year, city, school district, number of online course(s) taken, number of online course(s) passed, number of online course(s) failed, and final grade average 3. Student Information: Number of courses taken, Number of courses failed, Number of courses passed, Average individual student pass rate for all courses in academic year 2009-2010	1.Classification 2.Clustering	1. e-Learning System Logs 2. Student Information System Data 3. Course Evaluations Surveys

RP28	Students Previous Grades & Class Performance	• 1st written assignment • 2nd written assignment • 3rd written assignment • 4th written assignment	Classification	Student Information System Data
RP33	Students e-Learning activity	• Number of pieces of coursework done by the user in the course. • Total time in seconds that the user has taken in the assignment section. • Number of messages sent by the user in the forum. • Number of messages read by the user in the forum. • Total time in seconds that the user has taken in the forum section. • Number of quizzes seen by the user • Number of quizzes passed by the user • Number of quizzes failed by the user • Total time in seconds that the user has taken in the quiz section	Classification	e-Learning System Logs
RP34	Students e-Learning activity	• Number of messages written by the student • Number of words written by the student • Average score on the instructor's evaluation of the student's messages • Degree centrality of the student • Degree prestige of the student	1.Classification 2.Clustering	e-Learning System Logs
RP35	Course Evaluations	• Program workload • Program organization of teaching • Keep scheduled hours • Clear exam rules • Availability of lecturer outside class	Classification	Course Evaluations Surveys

		• Student's previous knowledge of the topic • Teacher ability to motivate • Clarity of teaching • Availability of lecturer inside class • On schedule with program • Workload–credit ratio • Prescribed reading list • Adequacy of lecture hall • Student interest in topic • Overall class satisfaction		
RP36	Students e-Learning activity	• Student access time series • Number of clicks per course.	Classification	e-Learning System Logs
RP43	1. Students Previous Grades & Class Performance 2. Students Demographics	• 4th Written assignment • 3rd Written assignment	Classification	Student Information System Data
RP44	Students e-Learning activity	• Total time online (sec) • Number of Course material viewed (by material category) (sec) • Average time per session (sec) • Total time material viewed (sec) • Number of Course material viewed • # Course material viewed (by material category) / Number of Course material released to date • Number of Logins • Average time material viewed (sec)	Classification	e-Learning System Logs

		• Total time course material viewed (by material category) (sec)		
RP67	1. Students Demographics 2. Students Experience Data	• Online learning readiness • Previous Online Experience • Gender • Online technologies self-efficacy • Age • Prior knowledge	Classification	Surveys
RP69	1. Students Demographics 2. Students Previous Grades & Class Performance	• Gender • High School Grade • Major in high school • Previous GPA • Number of Courses Registered • Sponsor • Advisory visit • English score • Attendance • Core Vs elective • Study time • Performance	Classification	Surveys
RP75	1. Students Demographics 2. Students Previous Grades & Class Performance 3. Students Social Data	• Gender • High School Grade • Mother Occupation Status • Discount	Classification	Surveys
RP81	1. Students Demographics	• Previous Semester Marks • Family Annual Income	Classification	Not reported

	2. Students Previous Grades & Class Performance 3. Students Social Data	• Student Category • Family Size • Attendance • High School Grade • Assignment Performance		
RP94	Students e-Learning activity	• Interactions with peers • Number of exercises performed • Number of forum posts	Classification	e-Learning System Logs
RP113	Students e-Learning activity	• Total # discussion messages posted • Total number of online sessions • Total time online • # Files viewed • # Assessments finished • # Assessments started • # Reply discussion messages posted • # Mail messages sent • # Assignments submitted • # Discussion messages read • # Web links viewed • # New discussion messages posted • # Mail messages read	Classification	e-Learning System Logs
RP120	Students Previous Grades & Class Performance	• Previous Semester Marks • Class Test Grade • Attendance • Assignment • Lab Work	Classification	Student Information System Data
RP123	1. Students Demographics 2. Students	• Midterm marks • Lab Test Grades • Students Practice	Classification	Student Information System Data

	Previous Grades & Class Performance	• Homework • Seminar Performance • High School Branch • Attendance		
RP127	Students Demographics	• Gender • Language medium • Stream of bachelor degree • Division obtained	Classification	Student Information System Data
RP128	1. Students Demographics 2. Students Previous Grades & Class Performance 3. Students Social Data	• Students grade in Senior Secondary education • Living Location • Medium of Teaching • Mother's Qualification • Students other habit • Family annual income status • Students family status	Classification	Surveys
RP142	1. Students Demographics 2. Students Previous Grades & Class Performance 3. Students Social Data	• Father's education level • Mother's education level • Academic performance rate • Average round • Mode round • Access year	Classification	Surveys
RP174	Students Social Data	• Number of Records on Each Category of Websites • Duration of Watching Online Videos • Students' Grades on Advanced Mathematics	Classification	Network Access Logs

RP198	Students e-Learning activity	• Total time spent on practical tasks • The time taken to hand in the task since the task was made available in the LE • Number of words in forum posts	Clustering	e-Learning System Logs
RP220	Students e-Learning activity	• Assignment upload • Forum add post • Forum update post • Forum view discussion • Assignment view • Assignment view all • Course view • Forum view forum • Resource view	Classification	e-Learning System Logs
RP277	Students e-Learning activity	• Course identification number • Number of assignments done • Number of quizzes passed • Number of quizzes failed • Number of messages send to forum • Number of messages read on the forums • Total time used on assignments • Total time used on quizzes • Total time used on forum	Classification	e-Learning System Logs
RP323	Students Previous Grades & Class Performance	• Cumulative GPA • Statistics grade • Calculus I grade • Calculus II grade • Physics grade • Dynamics mid-exam #1 score • Dynamics mid-exam #2 score • Dynamics mid-exam #3 score	Classification	Student Information System Data

Table 7: Summary of factors influencing students' performance and used data mining approaches and techniques

2.3.5. Synthesize the data

We extracted 215 distinct significant factors from the 34 research papers that affects the performance of students in their education life. Furthermore, during the data extraction process, we identified the category for each set of factors that were collected in each paper. As a result, we found 9 factor categories that the 215 factors belong to. The 9 categories of factors are described in Table 8.

Category	Description	Papers	Number of Papers
Students e-Learning activity	The activity logs of students in e-Learning systems, such as, the number of logins, the number of assignments done, number of quizzes done, etc.	RP1, RP7, RP9, RP23, RP25, RP33, RP34, RP36, RP44, RP94, RP113, RP198, RP220, RP277	14
Students Previous Grades & Class Performance	The grades or other performance indicators of students in previous courses, semesters, or years.	RP2, RP5, RP10, RP16, RP22, RP28, RP43, RP69, RP75, RP81, RP120, RP123, RP128, RP142, RP323	15
Students Environment	The attributes of a student environment, such as, the type of school, the type of the classroom, class period, etc.	RP5	1

Students Demographics	The demographics data of a student, such as, gender, age, nationality, ethnicity, etc.	RP5, RP16, RP22, RP25, RP43, RP67, RP69, RP75, RP81, RP123, RP127, RP128, RP142	13
Instructor Attributes	Information about the instructor of the student and his/her evaluation results.	RP10, RP12	2
Course Attributes	Information about the course or module the student is taking, such as, length of the course, difficulty, etc.	RP10, RP12	2
Students Social Information	Information related to the student social life, like the number of friends, if s/he smokes or not, etc.	RP16, RP22, RP75, RP81, RP128, RP142, RP174	7
Course Evaluations	Data collected from course evaluation surveys, such as, questions related to the clarity of the course, the level of satisfaction, etc.	RP25, RP35	2
Students Experience Information	Information about the experience of students about the course, such as the readiness of the student, and self-efficacy.	RP67	1

Table 8: Description of factors' categories

2.4. Phase 3: SLR Results

In this section, we report the results of our SLR research, where we will answer our research questions, and elaborate on the interesting results we came up with from the extracted data.

2.4.1. Distribution of research by factors' categories

We classified each article under one or more categories as described in Table 8. They are: (1) Students e-Learning activity, (2) Students previous grades & class performance, (3) Students environment, (4) Students demographics, (5) Instructor attributes, (6) Course attributes, (7) Students social information, (8) Course evaluations, and (9) Students experience information. As seen in Figure 2, the most common and widely used factor categories for predicting students' performance in higher education are students' previous grades & class performance (26%), followed by students' e-Learning activity (25%), students' demographics (23%), and finally, students' social information (12%). These 4 categories were presented in 86% of the research studies.

This finding is in agreement with the findings of a prior systematic literature review conducted by (Shahiri et al., 2015), which showed that the CGPA and internal assessment marks are the most frequently used attributes in the EDM community for predicting the students' performance; this matches our top factor category which represents the "students' previous grades & class performance".

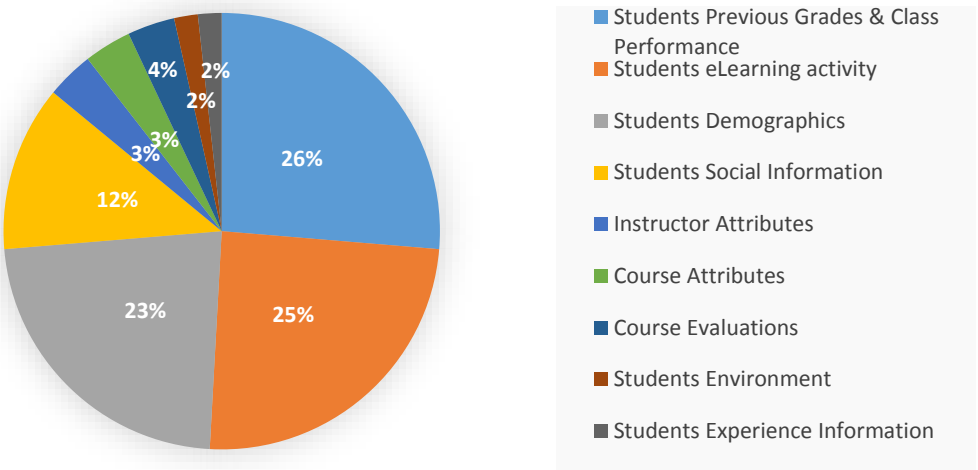

Figure 2: Distribution of research by category

The other 5 categories, which represent a total of 14% altogether, were used by few studies and did not appear often in other research articles; therefore, they were considered as ad hoc factors.

2.4.2. Distribution of research by year

Table 9: Distribution of research by year indicates that educational data mining research field was most popular on 2016, where more than 17.5% of the research was conducted on this year. And a significant increase of interest was started on 2012.

Year	Papers	Number of Papers
2009	RP142	1
2010	RP28, RP35, RP113	3
2011	RP120, RP127, RP128	3
2012	RP23, RP25, RP33, RP43	4
2013	RP16, RP34, RP323	3
2014	RP9, RP44, RP67, RP123	4
2015	RP1, RP81, RP174, RP220	4
2016	RP10, RP12, RP22, RP36, RP75, RP198	6
2017	RP2, RP7, RP69, RP94	4
2018	RP5, RP277	2

Table 9: Distribution of research by year

2.4.3. Distribution of research by data collection technique

Another dimension was added to the data collection, which is the techniques of the data collection. In fact, we identified five techniques of data collection, namely: (1) e-Learning system Logs, (2) Student Information System data, (3) Surveys, (4) Course evaluations surveys, and (5) Network access logs. Table 10: Summary of data types used by researchers summarizes the data collection techniques used by research papers in our data set.

Data collection technique	Description	Papers	Number of papers
e-Learning system logs	Logs obtained from e-Learning systems	RP1, RP7, RP9, RP23, RP25, RP33, RP34, RP36, RP44, RP94, RP113, RP198, RP220, RP277	14
Student Information System Data	Extracted data from student information systems, such as demographic data, admissions data, grades, etc…	RP2, RP5, RP16, RP22, RP25, RP28, RP43, RP120, RP123, RP127, RP323	11
Surveys	General survey obtained from students directly	RP16, RP22, RP67, RP69, RP75, RP128, RP142	7
Course Evaluations Surveys	Answers of course evaluation surveys, generally obtained at the end of each course.	RP10, RP12, RP25, RP35	4
Network Access Logs	Network dumb logs of students' activity on the internet and the university network.	RP174	1

Table 10: Summary of data types used by researchers

Figure 3: Distribution of research by data collection techniques illustrates the distribution of research articles by data collection techniques. As seen from Table 10: Summary of data types used by researchers and Figure 3, the most common data collection technique is the e-Learning system logs with nearly 38% of the research in the dataset. The second mostly used data collection technique is the student information systems data (30%), followed by surveys (19%), course evaluation surveys (11%), and network access logs (2%), respectively.

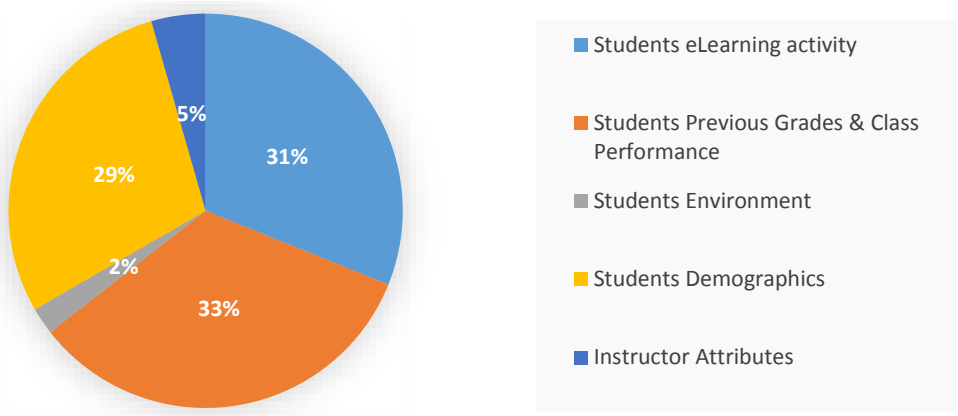

Figure 3: Distribution of research by data collection techniques

2.4.4. Distribution of research by data mining approaches

The main data mining approaches used in most of the studies are: (1) classification and (2) clustering. Table 11 summarizes the distribution of research in our data set by the two data mining approaches described. The main data mining approach is classification. It was found that all the research papers in the data set have used the classification approach to classify and predict the students' performance. On the other hand, only 4 research papers have used the clustering along with classification which was useful in order to find out how many different groups of students available in the data set and extract specific features of each group. This finding is in line with the findings of a prior study carried out by (Peña-Ayala, 2014), where it showed that the classification and clustering were the most typical data mining techniques used by EDM research.

Data mining approach	papers	Number of papers
Classifications	All	34
Clustering	RP2, RP23, RP25, RP34	4

Table 11: Distribution of research by data mining approaches

Furthermore, we extracted 141 data mining techniques/algorithms used by the 34 papers in our data set. Out of which, 74 were distinct. The algorithms are: 1NN, 3NN, ADTree, Apriori Algorithm, Artificial Neural Networks, BayesNet, Bivariate Regression, BP,

C4.5 Decision Tree, CART Decision Tree, CHAID, CitationKNN, Clustering, CRT Decision Tree, Decision Tree (DT), DecisionStump, DTNB, EM, FarthestFirst, Feed-Forward Neural Network (FFNN), G3P-MI, GP-ICRM, Gradient Boosting (GBM), HierarchicalClusterer, IBk, ICRM v1, ICRM v2, ICRM v3, ICRM2, ID3 Decision Tree, J48, Jrip, K-Means Clustering, K-Nearest Neighbour (k-NN), LADTree, LGR, Locally weighted linear regression, Logistic Regression, MILR, MLP (Multi-layer Perceptron) neural network, Model Trees, Multi-logistic Regression (MLR), Naïve Bayes, NaiveBayesSimple, Neural Network (NN), NLPCA, Nnge, OneR, PART, Prism, Probabilistic ensemble SFAM classifier (PESFAM), Probabilistic Ensemble Simplified Fuzzy ARTMAP, Proportional Odd Model (POM), Radial Basis Function (RBF) Network, Random Forest, Random Tree, RBF Network, Regression, Regression neural network model (RNN), REPTree, Resonance Theory Mapping (PESFAM), Ridor, RIPPER, Rule Induction, sIB, SimpleCart, SimpleKMeans, SMO, Support Vector Machine (SVM), Support Vector Ordinal Regression (SVOR), System for Educational Data Mining (SEDM), Visualization, WINNOW, Xmeans. However, the most commonly used algorithms that were used in 4 or more research papers (in more than 10% of the papers), are shown in Table 12.

Algorithm	**Frequency**	**Percentage (From the total number of papers (34))**
Naïve Bayes	13	38.2%
Support Vector Machine (SVM)	8	23.5%
Logistic Regression	6	17.6%
K-Nearest Neighbor (k-NN)	5	14.7%
ID3 Decision Tree	4	11.8%
C4.5 Decision Tree	4	11.8%
Decision Tree (DT)	4	11.8%

MLP (Multi-layer Perceptron) neural network	4	11.8%
Neural Network (NN)	4	11.8%

Table 12: Most commonly used data mining algorithms

Furthermore, we merged similar algorithms together into one category, for example, ID3 and C4.5 are both decision trees, so we grouped them together under the decision tree category. After doing so, we ended up with 7 categories of algorithms. Table 13 and Figure 3 shows the 7 groups of algorithms and the frequency of their usage in the research papers in our data set.

Algorithm	**Frequency**	**Percentage (From the total number of Algorithms (141))**
Decision Trees	35	24.8%
Naïve Bayes	14	9.9%
Artificial Neural Networks	13	9.2%
Regression	12	8.5%
Support Vector Machine	9	6.4%
K-Nearest Neighbor	8	5.7%
K-Means	3	2.1%
Other algorithms	47	33.3%

Table 13: Most commonly used algorithms by category

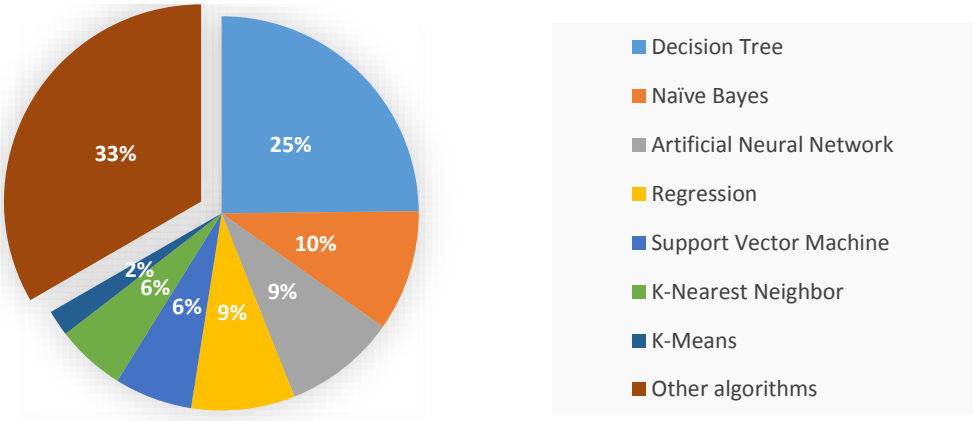

Figure 4: Most commonly used algorithms by category

As observed in Table 13 and Figure 4, the most commonly used category of data mining algorithms are Decision Trees, Naïve Bayes, and Artificial Neural Networks. This finding comes in line with the findings of prior systematic literature review studies (Peña-Ayala, 2014; Shahiri et al., 2015), where it was found that decision trees and Naïve Bayes are the most frequent data mining techniques used among EDM research.

3. Chapter Three: Research Methodology

This chapter provides a data mining implementation that is keen to apply most commonly used data mining techniques as per the findings of the systematic literature review in the previous chapter. The data mining implementation will be applied on a dataset extracted from a student information system of a local university in the United Arab Emirates. The extracted dataset's attributes is also compliant with the findings of the SLR in the previous chapter and other SLR researches as well. Finally, the last step of this DM implementation is analyze the produced outcomes and findings, and compare it with the literature. Figure

3.1. Dataset

This section describes the new dataset of the present study. Besides, a description of the data collection method and the data pre-processing is provided.

3.1.1. Data collection

The data in this study was collected from a local private university in the UAE. The university has more than 7000 active students and more than 35,000 graduates over the past 30 years. The data was extracted directly from the university database; a Microsoft SQL Server relational database. The data was extracted with a direct SQL query to the database using Microsoft SQL Server Management Studio. Taking into consideration all the joins and unions to get the data from the right tables and views. The dataset was limited to the past 5 years, from 2013 to 2018 including summer semesters, as well as, it was only limited to regular academic courses, in essence, any special courses such as training, orientation, studios, etc. were not considered. The final dataset size extracted from the database was 231,782 records of students' academic history.

3.1.2. Description of data attributes

The dataset of this study has 34 attributes in total within four types (Demographics, Course and Instructor Information, Student General Information, and Student Previous Performance Information). In this section, a detailed description of these attributes are provided. Table 14 provides a brief description of each attribute used in the study.

Attribute Type	Attribute	Attribute Description	Possible Extracted Values
Demographics	Student Nationality	The nationality of student	List of countries
	Gender	The gender of the student	Male, Female
	Age	A numeric value of the age of the student	A numeric range
	Student Program	The student program in the university	List of programs
	High School Name	The high school name where the student studied	List of high schools
	High School Country	The country where the student studied his high school	List of countries
	High School Branch	The branch of the student's high school	Scientific, Literature
	High School Curriculum System	The curriculum system of the high school	National High School, American, British, Indian, etc…
Course and Instructor Information	Academic Term	The term of which the course was taken	Fall, Winter, Summer
	Course Name	The course name the student is currently or was studying	List of course names
	Course Credit Hours	The number of credit hours of the course the student is currently or was studying	A numeric range
	Offering College	The college that are offering the course	List of colleges
	Instructor Position	The job position/rank of the instructor of the course	Assistant Professor, Associate Professor, Professor, etc…
	Instructor Gender	The gender of the instructor	Male, Female

	Instructor Nationality	The nationality of the instructor	List of countries
	Student & Instructor Genders	A comparison between the student gender and the instructor gender	If genders are the same, then “True”, otherwise, “False”
	Student & Instructor Nationalities	A comparison between the student nationality and the instructor nationality	If nationalities are the same, then “True”, otherwise, “False”
	Has Prerequisite	Whether or not the course has prerequisite	Yes, No
	Section Size	A numerical value of the size of the section (number of students in the class)	A numeric range
	Section Size Nominal*	A defined category based on the section size value	Small, Average, Large, Huge
	Class Timing*	The timings of the class	Morning, Evening, Mixed
Student General Information	Resident in Hostel	Whether or not the student is a resident in the university hostel	Yes, No
	Has Discount	Whether or not the student gets a discount from the university	Yes, No
	Is Sponsored	Whether or not the student is sponsored by a third-party contractor	Yes, No
	Number of Siblings in the University	Number of siblings that are or were studying in the university	A numeric range
Student Previous Performance Information	Attendance Warning	How many warnings the student received for the current course	1, 2, 3
	Number of Absences	The number of times the student was absent from the class	A numeric range between 1-100
	High School Percentage*	The percentage of the high school mark	A numeric range between 60 and 100
	High School Merit	A defined merit based on the high school percentage	Low Performer, High Performer, Excellent

	Math Average*	The average of the total marks of a student in all mathematics subjects s/he studied in the university	Failure, Low Performer, High Performer
	Physics Average*	The average of the total marks of a student in all physics subjects s/he studied in the university	Failure, Low Performer, High Performer
	University Requirements Average*	The average of the total marks of a student in all university requirements subjects (except Math and Physics) s/he studied in the university	Failure, Low Performer, High Performer
	Prerequisite Average*	The average of the total marks of a student in all the prerequisites of the subject s/he studying	Failure, Low Performer, High Performer
Class (Label)	Class (Label)*	The final class that the student has achieved. Calculated based on the student's final marks for the course s/he's studying	Failure, Low Performer, High Performer

Table 14: Description of data attributes

* Calculated attributes are described in detail in the next section (3.2.3. Data Preprocessing)

Furthermore, it is equally important to provide some data visualization in this section. Henceforth, Figure 5 provides a pie chart of the label attribute distribution. It is clear that the data is imbalanced with regards to class/label. Where most of the examples have a class of "High Performer" with 175,914 examples. Comparatively, the other two classes have very less number of examples, "Low Performer" having 37,020 examples, and "Failure" having 18,848 examples only. As a result, a provisioning procedure was taken into account in order to balance the data and avoid the accuracy paradox. Section 4.3 describes this procedure in detail.

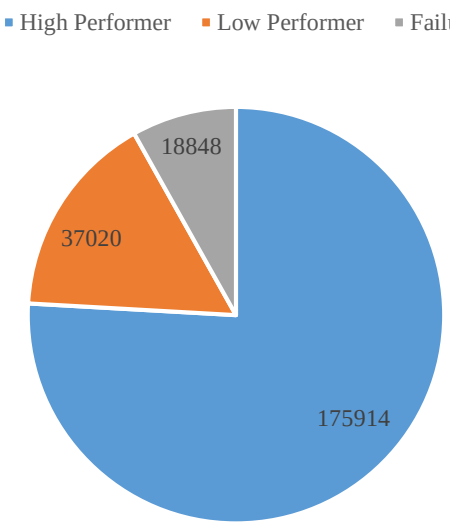

Figure 5: Class Distribution

3.1.3. Data Preprocessing

Some data preprocessing steps were performed on the data, which were split in two phases. This subsection describes these two phases and the related processes in detail.

3.1.3.1 Phase 1

The first phase of data preprocessing was done during the data extract from the database by the SQL query. Some attributes were generated based on other fields, as follows:

- Age: it was generated based on the difference between the student date of birth and the start date of the semester of the record
- High School Merit: it was generated based on the student high school marks. In detail:
    - Low Performer: was labeled for students with high school mark less than 75%
    - High Performer: was labeled for students with high school mark between 75% and 85%

    - Excellent: was labeled for students with high school mark greater than, or equal 85%

- Student and Instructor Genders: it was generated by comparing the student and instructor gender. If they were the same, it will be “True”, otherwise, “False”.
- Student and Instructor Nationalities: it was generated by comparing the student and instructor nationalities. If they were the same, it will be “True”, otherwise, “False”.
- Section Size Nominal: this attribute is based on the “Section Size” attribute, where:
    - Small: was given to classes of size less than 20 students
    - Average: was given to classes of size 20 to 39 students
    - Large: was given to classes of size 40 to 64 students
    - Huge: was given to classes of size greater than, or equal 65 students
- Class Timing: was calculated based on the class start and end times, where:
    - Morning: was given to classes starting from early morning, and ending no later than 1:30 PM.
    - Evening: was given to classes starting on 1:30 PM or later and ending after any time after that
    - Mixed: was given to classes starting in the morning and ending after 1:30 PM
- Math Average: is the average of the accumulated marks of a student in all mathematics subjects s/he studied in the university, where:

- Failure: is for averages less than 60 mark.
- Low Performer: is for averages greater than or equal to 60 and less than 75 mark
- High Performer: is for averages greater than or equal to 75 mark

- Physics Average: is the average of the accumulated marks of a student in all physics subjects s/he studied in the university, where:
  - Failure: is for averages less than 60 mark.
  - Low Performer: is for averages greater than or equal to 60 and less than 75 mark
  - High Performer: is for averages greater than or equal to 75 mark
- Prerequisite Average: is the average of the accumulated marks of a student in all the prerequisites of the subject, in which, s/he is studying, where:
  - Failure: is for averages less than 60 mark.
  - Low Performer: is for averages greater than or equal to 60 and less than 75 mark
  - High Performer: is for averages greater than or equal to 75 mark
- University Requirements Average: is the average of the accumulated marks of a student in all university requirements subjects s/he studied in the university, where:
  - Failure: is for averages less than 60 mark.
  - Low Performer: is for averages greater than or equal to 60 and less than 75 mark
  - High Performer: is for averages greater than or equal to 75 mark

- Class (label): is also calculated based on the student final marks for the course s/he's studying, where:
    - Failure: is for marks less than 60
    - Low Performer: is for marks greater than or equal to 60 and less than 75
    - High Performer: is for marks greater than or equal to 75

Important to realize, that "Math Average", "Physics Average", and "University Requirements Average" is calculated by excluding the course of the current record in case it is part of the same course collection. To clarify, let's assume that the current record course is MATH 101, and the collection of math courses taken by the student are MATH 101, 102, and 103. The "Math Average" of the current record is going to be the average of the marks of MATH 102 and 103 only without including the marks of MATH 101, otherwise, it will be considered cheating. Identically, the same applies to "Physics Average" and "University Requirements Average".

3.1.3.2 Phase 2

Second phase was done on RapidMiner software, where the actual data mining was done. This study uses RapidMiner 9 software for all its data mining tasks, as well as, for the second phase preprocessing tasks in this section.

As mentioned earlier, it is clear that the dataset used in this study is imbalanced, where most examples are of class "High Performer". This results in misleading performance measures, especially accuracy. Consequently, it was decided to use a balanced sample of the data, where all the class examples are of equal size. Therefore, we used under-sampling to achieve this goal, where the sample size of each class was downsized to match the size of the least class dataset. As a result, the total sample size was (18,848 × 3) 56,544, which is still good enough for a classification data mining study.

3.2. Data Mining Implementation

This section provides the data mining implementation plan and steps. All data mining tasks were carried out using RapidMiner 9 software, with an educational license.

3.2.1. Selecting data mining approaches and algorithms

Based on the findings of our systematic literature review in chapter two, it was found that the most frequently used data mining approach in the educational data mining research was classification. Additionally, the most frequently used data mining algorithms by category was as stated earlier, Decision Trees (DT), Naive Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Regression. Table 15 represents the data mining algorithms selected and applied in this study accordingly.

Algorithm Category	Algorithm Name	Abbreviation
Decision Tree	Decision Tree (RapidMiner's default)	DT
Decision Tree	Random Forest	RF
Decision Tree	Gradient Boosted Trees	GBT
Artificial Neural Networks	Deep Learning	DL
Naïve Bayes	Naïve Bayes	NB
Regression	Logistic Regression	LR
Regression	Generalized Linear Model	GLM

Table 15: Selected Data Mining Algorithms

Unfortunately, we could not use Support Vector Machines (SVM) algorithms because it requires only numerical attributes, whereas our data has a mixture of nominal and numerical attributes.

3.2.2. Data Mining Implementation

Before finalizing the data mining models, it is crucial to test multiple settings and parameters for each predictive algorithm. Similarly, we tested our data mining models with many combinations of settings and parameters before finalizing it to the

configurations shown in Table 16, keeping in mind that it only mentions changes to default values.

Algorithm	Settings/Parameters				
DT	Criterion: Information Gain	Confidence: 0.05	Minimal gain: 0.001	Minimal leaf size: 1	
RF	Criterion: Information Gain	Confidence: 0.05	Minimal gain: 0.001	Minimal leaf size: 1	Number of Trees: 50
GBT	Learning Rate: 1.0				
DL	No change				
NB					
LR					
GLM					

Table 16: Setting/Parameters for each algorithm

Furthermore, the performance validation were generated using a 10-fold cross validation. Where the data are split into ten folds, so the training is done using nine folds and testing with the remaining one-fold. The resulted accuracy is the average of the accuracies of the ten runs. In addition, we also used the Precision and Recall as performance measures in order to get a broader analysis of the performance of the algorithms.

Figure 6 shows the data mining processes in Rapid Miner.

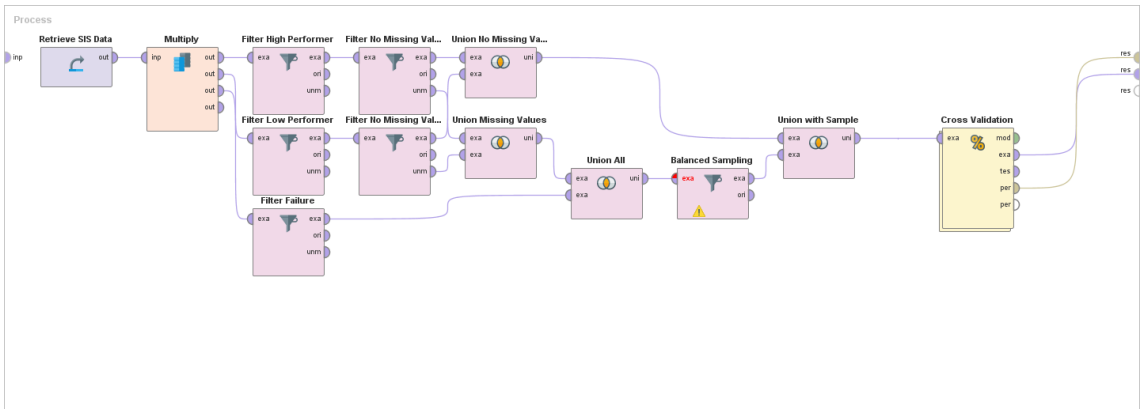

Figure 6: Rapid Miner Processes

4. Chapter Four: Results & Research Questions Answers

This chapter provides summarized answers to the research questions imposed in this dissertation. As well as, references to the corresponding sections in the dissertation where the full answer exists.

4.1. Results

Table 17 summarizes the performance of all data mining algorithms executed on our dataset as part of our data mining implementation.

Algorithm	Accuracy %	Recall %			Precision %		
		High Performer	Low Performer	Failure	High Performer	Low Performer	Failure
DT	64.76	59.62	60.01	74.75	63.38	53.08	80.42
RF	71.80	71.56	61.75	82.13	69.80	61.64	84.46
GBT	71.04	69.83	62.33	81.00	69.94	60.30	83.69
DL	51.86	12.43	98.05	45.67	87.23	40.72	99.08
NB	66.97	66.66	57.70	76.58	67.46	56.50	77.29
LR	71.76	71.98	62.62	80.71	69.85	61.71	84.59
GLM	61.21	79.05	34.30	70.02	59.54	52.09	69.41

Table 17: Performance of the seven data mining algorithms

As it can be seen in Table 17, Random Forest (RF) algorithm had the highest marginal accuracy score over the other algorithms; however, it does not have the best scores in all class Recalls and Precisions. RF had the best Recall of predicting “Failure” class. This indicates that RF is the best at predicting “Failure” class than other algorithms. However, these scores are very marginal compared to other algorithms, such as LR and GBT, where their scores is very close to each other.

On the other hand, a surprising result is noticed in Deep Learning (DL) results, were it was powerful enough to predict 98% of "Low Performer" class records, but very weak in predicting the other two classes, even though, the data was evenly balanced across the three classes. Nevertheless, Deep Learning is still a very powerful data mining technique/algorithm, and has a high potential to perform better in other implementations or data mining researches.

Another interesting observation is that almost all the algorithms had a very good Recall for "Failure" class over the other two classes, even though, at the begging, "Failure" was the least class in the dataset before under-sampling the data. This result shows the power of successfully balancing a dataset. However, we tried increasing the sample size of the other two classes slightly, in hope that their Recall might increase, but we did not notice any significant improvements in the Recalls nor the overall Accuracies.

Furthermore, it was noted that the least predicted class among the three classes was "Low Performer", where most of the algorithms faced difficulties predicting it. This might be reasoned to the fact that "Low Performer" is a middle class between "Failure" and "High Performer", where most of the noise usually exists.

Additionally, comparing the generated results with a baseline is usually practiced by researchers to emphasis their results. And since this is a newly created dataset, we chose the default model as our baseline, which is described as a model that predicts the class of all examples in a dataset as the class of its mode. This creates a baseline measure that we can compare our results with. In our case, since the dataset is balanced, and we have three classes; hence, the baseline Accuracy of the default model is easily calculated as 33.33%. Therefore, it is confirmed that the Accuracies of all the models generated in this study outperform the baseline model significantly.

Finally, we can answer our third research question by stating that the most appropriate data-mining algorithm for predicting students' academic performance using the data extracted from a student information system is Random Forest (RF).

Moving on to the second research question analysis and results. Table 18 shows the top 15 most important attributes according to information gain.

Attribute Type	Attribute	Importance
Demographics	High School Name	0.181
Student Previous Performance Information	University Requirements Average	0.162
Student Previous Performance Information	Math Average	0.144
Student Previous Performance Information	Physics Average	0.129
Student Previous Performance Information	Prerequisite Average	0.122
Student Previous Performance Information	Attendance Warning	0.088
Course and Instructor Information	Course Name	0.087
Student Previous Performance Information	High School Merit	0.029
Demographics	Student Program	0.025
Student General Information	Has Discount	0.021
Demographics	Gender	0.017
Demographics	High School Country	0.017
Demographics	Student Nationality	0.016
Course and Instructor Information	Offering College	0.012
Course and Instructor Information	Instructor Nationality	0.010

Table 18: Most important attributes according to Information Gain

As it can be noticed from Table 18, the most repeated attribute type is "Student Previous Performance Information" having six records out of 15. This outcome was very much expected, and hence, it confirms that the most important attribute type that mainly predicts a student's academic performance is his/her previous grades and general performance. All the averages that was calculated from the student previous performance in other subjects were very relevant to the student current performance. This finding is in line with

the finding of many prior researches (Asif et al., 2017; C. Anuradha Bharathiar, 2015; Fernandes et al., 2018; Gómez-Rey et al., 2016; S. B. Kotsiantis, 2012; Márquez-Vera et al., 2016, 2013; Shahiri et al., 2015).

Surprisingly, the first and most important attribute was "High School Name", which indicates that a student is going to perform very well or very bad in college based on the high school he attended. However, this attribute had a very large number of distinct values in the dataset, counting more than 3200 high schools. Nevertheless, it was still showing a very high relevance to the class.

Additionally, it was noted that there are four more demographical information that was identified as important in Table 18 other than "High School Name", specifically: "Student Program", "Gender", "High School Country", and "Student Nationality". This indicates that the demographical information holds an important value for predicting students' academic performance. This finding is inline with many prior research (Abazeed & Khder, 2017; Abu Saa, 2016; Araque et al., 2009; Badr El Din Ahmed & Sayed Elaraby, 2014; C. Anuradha Bharathiar, 2015; Fernandes et al., 2018; Inan, Yukselturk, & Grant, 2009; S. B. Kotsiantis, 2012; Márquez-Vera et al., 2016, 2013; Yukselturk et al., 2014). In detail, it is evident that the difficulty of a program affects students' academic performance, either negatively or positively. On the other hand, the student nationality and high school country were flagged as relevant to the prediction of students' academic performance, where students coming from certain countries performed better than others. Finally, it was evident by the prediction models that males and females have significant differences in their performance.

Comparatively, including the course names in the dataset seemed a bad idea at first, but after reaching out to the results, it was concluded that some courses in the university are harder than others, and some of them are easier. Hence, the prediction model confirms

this hypothesis, and uses this attribute effectively to predict students' performances accordingly. This also seems very important in order to warn students studying these courses in the future that they might face difficulties in getting higher grades, and provide them with opportunities to increase their efforts accordingly.

Finally, if the student have a discount from the university, it was noted that students' tend to perform better. This finding is in line with prior data mining research finding on the same subject (Abu Saa, 2016).

Given the above analysis and results, we can now answer our fourth research question. The most important attributes and the main predictors of students' academic performance among the selected attributes from a student information system is summarized in Table 18, and eventually, they belong to four types, Students' Demographics (Abazeed & Khder, 2017; Abu Saa, 2016; Araque et al., 2009; Badr El Din Ahmed & Sayed Elaraby, 2014; C. Anuradha Bharathiar, 2015; Fernandes et al., 2018; Inan et al., 2009; S. B. Kotsiantis, 2012; Márquez-Vera et al., 2016, 2013; Yukselturk et al., 2014), Student Previous Performance Information (Asif et al., 2017; C. Anuradha Bharathiar, 2015; Fernandes et al., 2018; Gómez-Rey et al., 2016; S. B. Kotsiantis, 2012; Márquez-Vera et al., 2016, 2013; Shahiri et al., 2015), Course and Instructor Information (Costantini et al., 2010; Gómez-Rey et al., 2016; Hung et al., 2012; Jiang et al., 2016), and some Student General Information (Abu Saa, 2016).

4.2. Discussion

This dissertation provided a comprehensive data mining research focused on predicting student academic performance and determining the main factors affecting a student academic performance. The study was designed to fully explore the educational data mining domain and was targeted to achieve interesting results related to the subject. The study contributed four fully answered research questions that are very useful to the EDM

community and future research on the same subject. The results were in line and agrees with most the related literature, and provides a broader idea of educational data mining in terms of students' academic performance prediction and the most influencing factors affecting students' academic performance. The study also provided a systematic literature review that reviewed 36 research papers related to the subject, and came up with results of distribution of research across multiple dimensions. The study overcame multiple challenges and limitations associated with the implementation of the research, where it lied in a challenging environment with many diversities.

4.3. Research Questions Answers

4.3.1 Research Question 1

Question: What are the most common and most frequently used factors affecting students' performance in higher education?

Answer: According to the results of the SLR study proposed in chapter 2, it was found that the most common and most frequently used factors affecting students' performance in higher education are students' previous grades & class performance (26%), followed by students' e-Learning activity (25%), students' demographics (23%), and finally, students' social information (12%). These findings were discussed in detail in section 2.4.1. Distribution of research by factors' categories and illustrated in Figure 2.

4.3.2. Research Question 2

Question: What are the most common and most frequently used data mining techniques used to analyze and predict students' academic performance?

Answer: According to the results of the SLR study proposed in chapter 2, it was found that the most common and most frequently used data mining techniques used to analyze and predict students' academic performance are Decision Trees, Naïve Bayes, and

Artificial Neural Networks. These findings were discussed in detail in section 2.4.4. Distribution of research by data mining approaches and illustrated in Figure 4.

4.3.3. Research Question 3

Question: What is the most appropriate data mining technique/algorithm that has the best results for predicting students' academic performance using real data extracted from a student information system?

Answer: According to the results of the data mining implementation study proposed in chapter 3, it was found that the most appropriate data mining technique/algorithm that has the best results for predicting students' academic performance using real data extracted from a student information system is Random Forest (RF). These findings were discussed in detail in section 4.1. Results and illustrated in Table 17 and Figure 2.

4.3.4. Research Question 4

Question: What are the main predictors of students' academic performance among the attributes selected from a student information system?

Answer: According to the results of the data mining implementation study proposed in chapter 3, it was found that the main predictors of students' academic performance among the attributes selected from a student information system are student's high school, university requirements grades, some specific courses, attendance warnings, number of absences, the course's prerequisites average, student program, high school grades and percentage, grades in math, having a discount, nationality, the course offering college, gender, and physics grades. These attributes belonged to four main types, students' demographics, student previous performance information, course and instructor information, and student general information. These findings were discussed in detail in section 4.1. Results and illustrated in Table 18.

5. Chapter Five: Conclusion and Future Prospects

This chapter provides the conclusion of this dissertation, as well as, the proposed future work.

5.1. Conclusion

This dissertation provided an educational data mining study that are focused to produce a standardized set of factors affecting students' academic performance. The study were split into two main parts. Part one, which is embodied in chapter 2, provided a systematic literature review that aimed to identify the most common and widely studied factors affecting students' academic performance in higher education by the EDM community. Additionally, it identified the most common data mining approaches, techniques, and algorithms used to classify and predict students' performance. We followed a systematic literature review methodology that consisted of multiple phases and steps. We started off by planning the review, from forming up the research questions, through setting up the inclusion and exclusion criteria, until deciding the data extraction strategy. Furthermore, the second phase consisted of the steps for conducting the review, kicked off by searching and identifying the research papers for the literature review, passing by assessing the quality of the selected research papers, and ended by extracting the data and synthesize it. Finally, we ended up by reporting the results of our SLR study, which discussed that the most common and widely used factors for predicting students' performance in higher education are students' previous grades & class performance, students' e-Learning activity, students' demographics, and students' social information. Additionally, results also showed that the most common and widely used data mining techniques in the educational data mining field are Decision Trees, Naïve Bayes, and Artificial Neural Networks.

The second part of the dissertation, embodied in chapter 3, provided a data mining research, and generated a model for student's performance prediction for data extracted from student information system. A gap was identified and addressed in this study, which was integrating data mining models and algorithms to student information systems' data. This is most helpful to educational institutions where it helps instructors and students to identify the weaknesses and factors affecting student's performance, and act as an early warning system to alert for predicted failures or low performance. Consequently, we were able to find out the most appropriate data mining algorithm to be used with student information systems that can predict students' academic performance accordingly. It was found that Random Forest (RF) is the most appropriate data mining technique used in this study. Furthermore, we were able to identify a list of most important and most relevant student attributes and their general types that have a direct effect to students' academic performance. Results showed that the most important students' attributes are the student's high school, university requirements grades, some specific courses, attendance warnings, number of absences, the course's prerequisites average, student program, high school grades and percentage, grades in math, having a discount, nationality, the course offering college, gender, and physics grades. These attributes belonged to four main types, students' demographics, student previous performance information, course and instructor information, and student general information. In fact, these findings supports and confirms the findings of the SLR study from the first part. Therefore, this set of attributes can be focused on in future EDM studies instead of wasting time extracting non-necessary information about students.

It was evident that the current study was limited to data extracted from student information systems only, however, due to this limitation, this study found a new gap between the world of educational data mining and student information systems' data, and addressed it accordingly. Of course, it would have been better to include data from other

student related systems, such as eLearning systems, to be able to harness even more results and conclusions through the analysis of such systems and find out their relationship with the student outcomes and achievements, however, this study focused on only one type of student systems to be more specific and concentrated. Furthermore, another limitation was the number of records that were accessible to this study was limited to the past five years only, where it would have been better to include even more historical records of students' data to enable the study to be more vast and broad.

5.2. Future Work

As a future work, researchers can benefit from the interesting outcomes of the systematic literature review proposed in this dissertation by employing it to their future research, particularly, the main results that suggests the most frequently used factors' categories affecting students' performance, as well as, the most frequently used data mining techniques. Not to mention, having a generic set of factors' categories provides limitless possibilities to tailor the use of these categories and come up with specific factors within the category for each educational institution, since it might differ from one place to another, and from time to time.

Finally, another possible future work includes connecting models produced by the data mining implementation of this dissertation to an institution's student information system to enable it to use the power of machine learning and data mining to warn students online once a possible failure or low performance is identified by the model.

6. References

Abazeed, A., & Khder, M. (2017). A Classification and Prediction Model for Student's Performance in University Level. *Journal of Computer Science, 13*, 228–233.

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology and Society*.

Abu Saa, A. (2016). Educational Data Mining & Students' Performance Prediction. *Internation Journal of Advamced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2016.070531

Al-Araibi, A. A. M., Mahrin, M. N. Bin, & Yusoff, R. C. M. (2016). A systematic literature review of technological factors for e-learning readiness in higher education. *Journal of Theoretical and Applied Information Technology*. https://doi.org/10.1177/1046496416665221

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers and Education*. https://doi.org/10.1016/j.compedu.2009.03.013

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*. https://doi.org/10.1016/j.compedu.2017.05.007

Badr El Din Ahmed, A., & Sayed Elaraby, I. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*. https://doi.org/10.13189/wjcat.2014.020203

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*. https://doi.org/10.1007/s10639-017-9616-z

Baradwaj, B., & Pal, S. (2012). Mining educational data to analyze student's performance. *Internation Journal of Advamced Computer Science and Applications*. https://doi.org/vol.2,No.6

Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security,*.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *Washington, DC: SRI International*. https://doi.org/10.2991/icaiees-13.2013.22

Budgen, D., & Brereton, P. (2006). Performing systematic literature reviews in software engineering. In *Proceeding of the 28th international conference on Software engineering - ICSE '06*. https://doi.org/10.1145/1134285.1134500

Burgos, C., Campanario, M. L., Peña, D. de la, Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*. https://doi.org/10.1016/j.compeleceng.2017.03.005

C. Anuradha Bharathiar, and T. V. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*. https://doi.org/10.17485/ijst/2015/v8i

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers and Education*. https://doi.org/10.1016/j.compedu.2016.02.006

Chamizo-Gonzalez, J., Cano-Montero, E. I., Urquia-Grande, E., & Muñoz-Colomina, C. I. (2015). Educational data mining for improving learning outcomes in teaching

accounting within higher education. *International Journal of Information & Learning Technology*. https://doi.org/10.1108/IJILT-08-2015-0020

Costantini, P., Linting, M., & Porzio, G. C. (2010). Mining performance data through nonlinear PCA with optimal scaling. *Applied Stochastic Models in Business and Industry*. https://doi.org/10.1002/asmb.771

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2018). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2018.02.012

Gamulin, J., Gamulin, O., & Kermek, D. (2016). Using Fourier coefficients in time series analysis for student performance prediction in blended learning environments. *Expert Systems*. https://doi.org/10.1111/exsy.12142

Gómez-Rey, P., Fernández-Navarro, F., & Barberà, E. (2016). Ordinal regression by a gravitational model in the field of educational data mining. *Expert Systems*. https://doi.org/10.1111/exsy.12138

Hasheminejad, .-H, & Sarvmili, M. (2018). S3PSO: Students' Performance Prediction Based on Particle Swarm Optimization. *Journal of AI and Data Mining*.

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.04.002

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*. https://doi.org/10.1016/j.compedu.2012.08.015

Hung, J., Hsu, Y.-C., & Rice, K. (2012). Integrating Data Mining in Program Evaluation

of K-12 Online Education. *Educational Technology & Society*. https://doi.org/10.1207/s15327752jpa8502

Inan, F. a, Yukselturk, E., & Grant, M. M. (2009). Profiling potential dropout students by individual characteristics in an online certificate program. *International Journal of Instructional Media*.

Jiang, Y. H., Javaad, S. S., & Golab, L. (2016). Data mining of undergraduate course evaluations. *Informatics in Education*. https://doi.org/10.15388/infedu.2016.05

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., Haydock, M., Kiron, B. D., … Haydock, M. (2012). Analytics: The Widening Divide. *MIT Sloan Management Review*.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Engineering*. https://doi.org/10.1145/1134285.1134500

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*. https://doi.org/10.1016/j.infsof.2008.09.009

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-011-9234-x

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2010.03.010

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. *Computers & Education*. https://doi.org/10.1016/j.compedu.2013.10.009

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*. https://doi.org/10.1016/j.compedu.2009.09.008

Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*. https://doi.org/10.1080/19439342.2012.711342

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*. https://doi.org/10.1111/exsy.12135

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*. https://doi.org/10.1007/s10489-012-0374-8

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in moodle learning management system: A case of Mbeya University of science and technology. *Electronic Journal of Information Systems in Developing Countries*. https://doi.org/10.1002/j.1681-4835.2017.tb00577.x

Pandey, U. K., & Pal, S. (2011). Data Mining : A prediction of performer or underperformer using classification. *(IJCSIT) International Journal of Computer Science and Information Technologies*.

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2013.08.042

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education*. https://doi.org/10.1016/j.compedu.2013.06.009

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. https://doi.org/10.1109/TSMCC.2010.2053532

Salloum, S. A., Al-Emran, M., Abdallah, S., & Shaalan, K. (2017). Analyzing the Arab Gulf Newspapers Using Text Mining Techniques. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 396–405).

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. In *Studies in Computational Intelligence* (Vol. 740). Springer. https://doi.org/10.1007/978-3-319-67056-0_18

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017). Mining Text in News Channels: A Case Study from Facebook. *International Journal of Information Technology and Language Studies*, *1*(1), 1–9.

Salloum, S. A., Mhamdi, C., Al-Emran, M., & Shaalan, K. (2017). Analysis and Classification of Arabic Newspapers' Facebook Pages using Text Mining Techniques. *International Journal of Information Technology and Language Studies*, *1*(2), 8–17.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. In *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2015.12.157

Wook, M., Yusof, Z. M., & Nazri, M. Z. A. (2017). Educational data mining acceptance among undergraduate students. *Education and Information Technologies*. https://doi.org/10.1007/s10639-016-9485-x

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.09.034

Yadav, S., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology & Creative Engineering*.

Yadav, S. K., & Pal, S. (2012). Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification. *World of Computer Science and Information Technology Journal WCSIT*. https://doi.org/10.1142/9789812771728_0012

Yukselturk, E., Ozekes, S., Türel, Y. K., Education, C., Ozekes, S., Türel, Y. K., & Education, C. (2014). Predicting Dropout Student : an Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and E- Learning*. https://doi.org/10.2478/eurodl-2014-0008

Zafra, A., & Ventura, S. (2012). Multi-instance genetic programming for predicting student performance in web based educational environments. *Applied Soft Computing*. https://doi.org/http://dx.doi.org/10.1016/j.asoc.2012.03.054

Zaza, S., & Al-Emran, M. (2016). Mining and exploration of credit cards data in UAE. In *Proceedings - 2015 5th International Conference on e-Learning, ECONF 2015*. https://doi.org/10.1109/ECONF.2015.57

Zhou, Q., Zheng, Y., & Mou, C. (2015). Predicting students' performance of an offline course from their online behaviors. In *2015 5th International Conference on Digital Information and Communication Technology and Its Applications, DICTAP 2015*. https://doi.org/10.1109/DICTAP.2015.7113173