PREDICTING STUDENTS' STEM ACADEMIC PERFORMANCE IN MALAYSIAN SECONDARY SCHOOLS USING EDUCATIONAL DATA MINING

By

MOHAMMAD IZZUAN BIN TERMEDI @ TERMIJI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Philosophy

January 2023

FPP 2023 4

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the Degree of Philosophy

PREDICTING STUDENTS' STEM ACADEMIC PERFORMANCE IN MALAYSIAN SECONDARY SCHOOLS USING EDUCATIONAL DATA MINING

By

MOHAMMAD IZZUAN BIN TERMEDI @ TERMIJI

January 2023

Chair : Aini Marina Ma'rof, PhD
Faculty : Educational Studies

Data mining has been widely applied in educational area recently and it is commonly called Educational Data Mining (EDM). Via data and success analysis, schools may recognize and improve learning plans for students who fail to fulfil their needs. It is necessary in the era of big data to treat data as an opportunity for schools to become a data-driven organization. However, predicting students' performance in Malaysian setting of diverse background and from huge dataset to inform secondary students' performance remain unclear. The aim of the study is primarily to define the input and environment variables that better predict students' progress and to establish a methodological approach using students' data which can be used by academics, schools and the educational ministry to assess students' performance.

Data mining is described in the CRISP-DM conceptual model in this research. CRISP is a cross-industry standard data mining process which consists of six stages of the normal lifecycle of a CRISP-DM project.

Design and development research (DDR) are chosen to be the conducted approach for this analysis. It proceeds through three phases of Need Analysis, Development of the Model and Evaluation of the Model. Four different data mining classification algorithms which are Random Forest, PART, J48 and Naive Bayes will be used on the dataset. This study is expected to investigate the process through utilize classification to help to predict students' performance. The target population for this study was all upper secondary students in Malaysia taking Science stream. Ten-fold cross validation uses a streamlined random sample technique to separate the entire dataset into 10 reciprocal sets. The data collected in this study are from two sources that consist of twenty-one attributes.

The first one is collected from the Education Repository in the Ministry of Education mainly APDM (Aplikasi Pangkalan Data Murid) and second from SAPS (Sistem Aplikasi Peperiksaan Sekolah). WEKA software will be used as a data mining software for all analyses.

Findings indicated that the high influential attributes in predicting students' academic performance are district, type of school, dual language program (DLP) status, religion, dormitory, nationality, guardian's job, guardian's salary group, and mid-year exam results. The results showed that each of the four classification algorithms have an average prediction accuracy of more than 70%, and J48 outperforms other classifiers based on accuracy and classifier errors.

This study has provided a methodological approach that will inform the ministry of education on how to predict students' performance using data available in the Education Data Repository. Then, this study was significant because the results of this study can help inform the ministry of education, teachers, and parents with valuable insight on key factors that can be used to help improve academic performance as well as students' success. Finally, this study can identify academically-at-risk students and develop early intervention.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MERAMAL PRESTASI AKADEMIK STEM PELAJAR DI SEKOLAH MENENGAH MALAYSIA MENGGUNAKAN PERLOMBONGAN DATA PENDIDIKAN

Oleh

MOHAMMAD IZZUAN BIN TERMEDI @ TERMIJI

Januari 2023

Pengerusi : Aini Marina Ma'rof, PhD
Fakulti : Pengajian Pendidikan

Perlombongan Data telah digunakan secara meluas dalam bidang pendidikan baru-baru ini dan ia biasanya dipanggil Perlombongan Data Pendidikan (EDM). Melalui data dan analisis kejayaan, sekolah mungkin mengiktiraf dan menambah baik rancangan pembelajaran untuk pelajar yang gagal memenuhi keperluan mereka. Adalah perlu dalam era data besar untuk menganggap data sebagai peluang untuk sekolah menjadi organisasi yang dipacu data. Walau bagaimanapun, meramalkan prestasi pelajar dalam latar belakang Malaysia yang pelbagai dan daripada set data yang besar untuk memaklumkan prestasi pelajar sekolah menengah masih tidak jelas. Matlamat kajian adalah terutamanya untuk menentukan pembolehubah input dan persekitaran yang lebih baik untuk meramalkan kemajuan pelajar dan untuk mewujudkan pendekatan metodologi menggunakan data pelajar yang boleh digunakan oleh ahli akademik, sekolah dan kementerian pendidikan untuk menilai prestasi pelajar.

Perlombongan data diterangkan dalam model konseptual CRISP-DM dalam penyelidikan ini. CRISP ialah proses perlombongan data standard merentas industri yang terdiri daripada enam peringkat kitaran hayat biasa projek CRISP-DM.

Penyelidikan reka bentuk dan pembangunan (DDR) dipilih sebagai pendekatan yang dijalankan untuk analisis ini. Ia diteruskan melalui tiga fasa iaitu Analisis Keperluan, Pembangunan Model dan Penilaian Model. Empat algoritma klasifikasi perlombongan data berbeza iaitu *Random Forest, PART, J48* dan *Naive Bayes* akan digunakan pada set data. Kajian ini diharapkan dapat menyiasat proses melalui penggunaan klasifikasi untuk membantu meramal prestasi pelajar. Populasi sasaran kajian ini adalah semua pelajar sekolah menengah atas di Malaysia yang mengambil aliran Sains.

Pengesahan silang sepuluh kali ganda menggunakan teknik sampel rawak diperkemas untuk memisahkan keseluruhan set data kepada 10 set timbal balik. Data yang dikumpul dalam kajian ini adalah daripada dua sumber yang terdiri daripada dua puluh satu atribut. Yang pertama dikumpul dari Repositori Pendidikan di Kementerian Pendidikan terutamanya APDM (Aplikasi Pangkalan Data Murid) dan kedua dari SAPS (Sistem Aplikasi Peperiksaan Sekolah). Perisian WEKA akan digunakan sebagai perisian perlombongan data untuk semua analisis.

Dapatan kajian menunjukkan bahawa atribut yang berpengaruh tinggi dalam meramal prestasi akademik pelajar ialah daerah, jenis sekolah, status program dwi bahasa (DLP), agama, asrama, kewarganegaraan, pekerjaan penjaga, kumpulan gaji penjaga, dan keputusan peperiksaan pertengahan tahun. Keputusan menunjukkan bahawa setiap empat algoritma pengelasan mempunyai purata ketepatan ramalan lebih daripada 70%, dan J48 mengatasi pengelas lain berdasarkan ketepatan dan ralat pengelas.

Kajian ini telah menyediakan pendekatan metodologi yang akan memaklumkan kepada kementerian pendidikan tentang cara meramal prestasi pelajar menggunakan data yang terdapat dalam Repositori Data Pendidikan. Kemudian, kajian ini adalah signifikan kerana hasil kajian ini dapat membantu memaklumkan kepada kementerian pendidikan, guru, dan ibu bapa dengan pandangan yang bernilai tentang faktor-faktor utama yang boleh digunakan untuk membantu meningkatkan prestasi akademik serta kejayaan pelajar. Akhir sekali, kajian ini dapat mengenal pasti pelajar berisiko akademik dan membangunkan intervensi awal.

ACKNOWLEDGEMENTS

Looking back on this dissertation journey, I am fascinated by the path I have taken, what I have experienced, and how many people have helped me along the way. The most valuable thing I learnt during this process is about the strength of determination. I learned that it would not actually take you to be the most knowledgeable or brightest person to receive a dissertation from a Ph.D. programme. You ought to stay concentrated on completing the project, and then making things happen and be willing to get the energy to get the job finished. Furthermore, I discovered that learning is not a product but a process. True learning needs the time that one can bring into it, and true learning has the struggle that one would possess in order to continue learning. As long as you don't stress, what you can do is just trust the mechanism and it will tip in your favor.

First, I would like to thank Allah, the most gracious and the most merciful, for granting me the courage and desire to write this thesis. Next, I would like to thank my parents who have been actively supporting my aspirations since I was born. I'm most grateful for you, Mom (Arbakyah binti Md Zin) and Dad (Termedi bin Zakaria) and my Mother-in-law (Masda binti Drawis), for encouraging me to continue with my graduate studies. I'm thankful to all my siblings for keeping confidence in me.

I would like to thank my committee members for all their encouragement and courage. Thanks to Dr. Aini Marina, Chair of my dissertation committee. Your help during three years of my study have really transformed this thesis into a real masterpiece. Thank you for your participation. Thank you for still being able to meet up when I needed you and also for encouraging me to keep moving forward and stick to the study phase. Associate Professor Dr. Habibah, you are a great tutor for us and still motivate us to do our utmost. You showed me how important becoming a successful researcher is in higher education. Thank you very much for agreeing to serve on my thesis team, Dr. Iskandar. You are among the coolest person I have ever met. Prof. Ismi, I want to thank you for what you have provided for my work and learning. Words cannot thank you enough for what you have done for me to get to where I am now.

I am extremely grateful to my beloved wife Nur Hazirah binti Baharin for she has been a great aid, always been there for me with words of motivation and have quietly been putting up with all my absences for the past years. Last but not least, I'd like to share my eternal and unconditional love for all my children, Izz Imanina, Izz Qayyum, Izz Muhammad Sufyan, Izz Huwayna, Izz Eylana, and Izz Hamza who have become an example and pillar of positive energy and power for me through tough times.

Lastly, I would like to thank you all for being a part of this dissertation. To 4IR Innovation Leaders, I want to thank you all for always caring about how my ideas are progressing, and for sending me words of encouragement. To my professors in the doctorate programme, thank you very much for your guidance within and outside the classroom, for the lessons you give of how to be a better educator, and for continuing to inspire me by becoming an outstanding scholar. To all my colleagues who are already in the Ph.D. program, thank you for always holding my back. Thanks for your companionship, encouragement and for your friendship. You guys are a wonderful and loving companion to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

BACKGROUND OF THE STUDY

1.1 Introduction

Technology entered the education system when we first hit the new millennium, and both students and teachers began clearly to integrate technology. We then noticed how technologies progressed, including widespread infiltration of a more user-generated internet (Chen et al., 2018). Students not only had their own links to information, the immersive learning process, but platforms to freely connect with other students. At the time, schooling was no longer deemed based on a back and forth between students and teachers, but instead followed a more networked method, with students providing their own direct connection to multiple sources of learning. This encouraged the creation of a more customized learning process to celebrate creativity and special study methodology. According to the World Economic Forum (2016), we are now in a modern period of the age named the Fourth Industrial Revolution (4IR).

The Fourth Industrial Revolution (4IR) centered on the application of advanced technologies, artificial intelligence and robotics. Primary and secondary schools, particularly teachers, must train their students for an environment in which these cyber-physical structures are widespread across all sectors to continue producing good students (Fullan et al., 2014). This ensures students to learn this technology as part of the program, adjust their approach, and use it to enhance school experience.

The Fourth Industrial Revolution (4IR) is becoming very crucial in the sense that it is not only will change the things we value and the way we value them but also in how we live, work, and communicate. In other words:

> We are at the beginning of a revolution that is fundamentally changing the way we live, work, and relate to one another. In its scale, scope and complexity, what I consider to be the fourth industrial revolution is unlike anything humankind has experienced before. (Schwab, 2017, p.9)

In order to catalyze innovation, creativity and job sustainability through education in industry 4.0, we need to invest in Science, Technology, Engineering & Mathematics (STEM) programs. STEM is a curriculum based on the idea of educating students in four specific disciplines — science, technology, engineering and mathematics — in an interdisciplinary and applied approach (Lee et al., 2019). Skills developed by students through STEM provide them with the foundation to succeed in school and beyond. STEM education will also encourage the increase in critical thinking skills which needed in the workforce to support the growth of the economy during the industrial revolution.

The School Management Division, Ministry of Education Malaysia has reported that there are 83,608 students registered in Form 4 Science stream in all upper secondary schools in Malaysia for the year 2018 (Ministry of Education Malaysia, 2019). Form 4 Science students definitely need to take STEM related subjects (Ali et al., 2021). This means that a very large scale of data or big data concerning all these students need to be stored. Consequently, it will also be a problem to analyze this big students' data in terms of the hidden knowledge inside them. The new age today belongs to science and technology, hence the need of educational technology. Educational technology studies the effect of science and technology upon education. In other words, science and technology are used under educational technology. This clearly shows that educational technology is in fact parallel with 4IR. One of the pillars among the nine pillars of 4IR is Big Data. Apparently, parallel with the fourth industrial revolution, analyzing such big data has made data mining an important role to play. Malaysia has recognised the importance of preparing its future workforce for the rapidly changing and technology-driven global economy, and as a result, STEM education has been gaining traction in the country's secondary school context in recent years.

The government of Malaysia has made efforts in recent years to expand access to STEM education and inspire more students to choose these subjects as a profession. Efforts in this direction have included the creation of new STEM curriculum, the installation of STEM infrastructure in schools, and the provision of funds for STEM-related study and investigation (Ministry of Education Malaysia, 2013). Not only have government agencies and nonprofits been pushing STEM education, but they have also been actively supporting schools and students. This has involved the creation of programmes and initiatives, as well as teaching materials and resources, with the goal of piquing students' interest in STEM topics and inspiring them to seek further education in these areas.

Data mining is one of the most popular techniques to analyze students' performance (Shahiri et al., 2015a). Data mining has been widely applied in educational area recently (Romero & Ventura, 2010) and it is commonly called Educational Data Mining (EDM). A report on the future of Educational Technology was commissioned by the United States National Science Foundation in 2009. The report focused on the role and impact of computing and technology in education, including research recommendations and a 2030 education vision. The report discusses seven areas of information and communications technologies which are likely to have a significant impact on learning and instruction. Essentially, Educational Data Mining (EDM) is one of the seven areas (Huang et al., 2019). Data mining is the process of extracting useful information from a huge amount of data. One of the most common applications of data mining is the use of different algorithms and tools to estimate future events based on previous experiences (Perner, 2015).

Schools are heading towards a more customised way of learning. Via data and success analysis, schools may recognise and improve learning plans for students who fail to fulfill their needs. 4IR, mainly in education, embraces this growth in analytics and uses it to consider each student separately and agree that all learning environments are specific and that the predicted results are specific. A vast "digital ocean" of student data is generated in classrooms, which can be useful observations if analysed. It is necessary in the era of big data to treat data as an opportunity for schools to become a data-driven organisation. The benefit is that schools have the resources to increase effectiveness, performance and shift decisions from perception to reality in order to ensure better and knowledgeable decision making (Mokhtar et al., 2019).

In Malaysia, the Education Data Repository was developed by the Ministry of Education to incorporate and connect the major Education Management Information System (EMIS) systems and applications. At the moment, 12 of the main programs and applications are included in the Repository. The systems are both separate, but the database allows them to share, interpret, and present data using a single information platform called the KPM Dashboard. The KPM Dashboard allows users at the ministry level to view and analyze data generated from the 12 applications. However, within the Ministry of Education, the KPM Dashboard is only available to top management and key personnel, including State Education Officers and District Education Officers, recognizing the Portal's ability to compare various data sources and to help provide insight into the education system. This is a key constraint on its use at the level of education (UNICEF, 2020). The Ministry of Education's ICT Transformation Plan 2019–2023 was developed in 2018 and supports the Education Data Repository. The plan proposes a single interconnected data sharing network for all structures and applications, providing a single gateway to the Ministry of Education's access to education data, and providing access to information at the national, divisional, district and school levels. The policy stresses intra-ministry coordination headed by the Information Technology Section. However, the project is not expressly supported, and the lack of financial resources limits the adoption of the initiative (Pacific, 2020).

1.2 Research Problem

Over the years, technologies have evolved continuously and drastically, and it also continue to change education. For example, all data involving students are written by teachers in lots of files. But now, as technology changed, that is when we are in the midst of 4IR, all students' data are stored online. All the data can be easily reachable within second. The Education Data Repository was created in 2017 with the goal of consolidating the processes data and initiating the process of streamlining data recording and reporting. 12 systems and 12 applications are included in the repository which enables broad data analytics, such as descriptive, diagnostic, predictive and prescriptive analytics, to be carried out simultaneously through data collected from these systems and applications in 2019. Analyzed data are provided through data analysis and tabular type through the Kementerian Pendidikan Malaysia (KPM) dashboard, which is used as guidance for decision-making and policymaking and for tracking compliance at central, state and district levels. However, beyond the district school office KPM Dashboard is not accessible and thus the usage of infographics and other data visualizations at school level is restricted (UNICEF, 2020).

As Huang et al. (2019) state, “Unfortunately, educational systems - primary, secondary, and postsecondary - have made limited use of the available data to improve teaching, learning, and learner success” (p.232). Due to the huge amounts of student information being gathered about academic achievement, coupled with the widespread usage of measuring and assessment points in students have reached in their academic path, educational study has been invaded by information that is either worthless or misused (Murtaugh et al., 1999). Despite the field of education lagging behind other sectors, there has been an explosion of interest in analytics as a solution for many current challenges, such as retention and learner support (Siemens, 2013). This is also applied for the case of students’ data in the Ministry of Education. Systems and applications were created independently and owned by separate departments, and there was a lack of coordinated attempt to integrate all these results. Often then served in silo, thus limiting the freedom to apply data to any study other than descriptive research (UNICEF, 2020).

A crucial topic of focus in education nowadays is predicting students' STEM success. STEM areas are essential for fostering innovation, economic expansion, and social advancement. To address global issues and promote breakthroughs across numerous sectors, it is crucial to ensure that the STEM workforce is highly qualified and competent. Identifying students who may be at risk of doing poorly or dropping out of STEM-related courses or programs is an important consideration when predicting students' success in STEM. Early identification of struggling learners enables teachers to act quickly with specialized assistance and interventions to enhance their academic performance and raise the possibility that they will graduate (Riegle-Crumb et al, 2012). Additionally, anticipating students' success in STEM subjects enables educators and institutions to modify their teaching methods and curricula to better suit students' requirements and preferences. As a result, learning is more effective and interesting, which helps students achieve better results and develop a greater interest in STEM fields (Alkan et al. 2019). To better meet the requirements of each student, educational institutions might use predictive models to distribute resources like money, qualified teachers, and technology. Institutions may maximize the effect of investments by concentrating on students who are likely to perform in STEM (Bienkowski et al. 2012).

Predictive models may also support the early identification and development of extraordinary STEM talent. The identification of exceptional STEM students may result in specialized enrichment programs, mentoring opportunities, and scholarships that support the growth of future innovators and specialists (Bybee & R. W., 2013). In addition, predicting STEM student performance guarantees that graduates are suitably equipped for the needs of the labor market. In an increasingly tech-driven world, a trained STEM workforce is essential for fostering economic development, innovation, and competitiveness (National Academy of Engineering, 2012).

Current research undeniably well predicts academic performance of students using data mining. But then, all the data mining techniques used to predict student performance were basically in their own specific domain. However, predicting students' performance in Malaysian setting of diverse background and from huge dataset to inform secondary students' performance in STEM remain unclear. This means that not all the data mining techniques might be suitable in predicting student performance. Despite many publications, including case studies, on educational data mining, it is still difficult for educators – especially if they are a novice to the field of data mining – to effectively apply these techniques to their specific academic problems (Alyahyan & Düştegör, 2020). Therefore, this study focusing on integrating different data mining techniques to build the best model to predict students' academic performance in Malaysian secondary schools.

As mentioned earlier, in order for students to survive in the future, that is getting a decent job, they need to be prepared to face 4IR. This specifically means that the need of STEM education to provide the students with all the necessary knowledge. Workforce related to STEM fields has become increasingly important. Sadly enough to say the policy of 60 to 40 percent of students for Science to Arts field at the tertiary education mandated by the Ministry of Education Malaysia since 1967 yet to be met (Halim, L.; Meerah, 2016). According to the Malaysian Examinations Syndicate (Lembaga Peperiksaan Malaysia), in 2017, only 27% of the candidates taking Physics and Chemistry, 21% Biology and 35% Additional Mathematics in the Sijil Pelajaran Malaysia (SPM), or the Malaysian Certificate of Education whereas in 2018, only 26% of the candidates taking Physics and Chemistry, 20% Biology and 34% Additional Mathematics. 4IR requires students to perform in STEM subjects but somehow students lose interest in the subjects. The problems of the decreasing enrollment of science students at secondary school level of Malaysian secondary students in international assessment studies point to a serious challenge for the ministry of education to improve STEM education. Therefore, there is a need to study the factors that lead to students' academic performance especially in STEM subjects in secondary schools.

One of the systems created in the EMIS is called School Examination Analysis System (SAPS). It is believed that SAPS is the most effective for teachers since, utilizing SAPS, schools are able to set targets for each student on the basis of their previous results (UNICEF, 2020). But is it enough just relying with the previous exam grade to predict the future achievements of a student? Again, the need to study the factors that lead to students' academic performance ought to be conducted. Many of the previous research well predict academic performance using data from higher education or university level. They also have identified factors that affects students' performance in general.

The researcher will locate the assessment of academic success among the activities where educational data plays a significant role and where literature has already made some progress. A few educational institutions are now informed that students will be able to promote higher levels of academic achievement through early inference of future academic success. This will lead to the design of differentiated actions addressing various classes of students according to their capacity and can also lead to a more effective distribution of resources from the institutions (Miguéis et al., 2018).

1.3 Research Objectives

The aim of the study is primarily to define the input and environment variables that better predict students' progress and to establish a methodological approach using students' data which can be used by academics, schools and the educational ministry to assess students' performance. This research will build and compare the statistical predictive data mining models like Random Forest, PART, J48 and Naive Bayes. All of these models will be improved to fit the data of students' performance and then evaluated to determine the best model for data mining. This research also considers significant characteristics of students who succeed in academic and non-excellent students. Furthermore, data taken from the Repository will be analyzed extensively through data mining process to ensure the quality of the data. Finally, the study will contribute in efficiency of the data mining techniques used in secondary schools and to help the ministry of education to make better use of data mining techniques in order to inform the performance of students in academics. Hence, the aim of this research is to predict academic performance among form 4 students in Malaysian Secondary Schools considering more accurate and credible students' academic data, using different data mining techniques.

Ultimately, the objectives of conducting this research are as follows:

1. To identify the needs for developing predictive models of students' STEM performance in Malaysian Secondary Schools using data mining.
2. To identify and select appropriate predictor variables or independent variables that can be used as the inputs of predictive models in predicting students' STEM performance.
3. To identify and select the best data mining techniques: Random Forest, PART, J48 and Naive Bayes for developing predictive models in predicting students' STEM performance.
4. To validate the developed models using the data collected in upper secondary schools in Malaysia and identify academically-at-risk students.

The model will produce useful and hidden knowledge to students, teachers and parents to take appropriate action to improve students' results and contribute to a better-quality education especially in STEM related subjects.

1.4 Research Questions

In order to address each research objectives of this study, the most important questions arise are as follows:

1. What are the trends of predicting students' performance using data mining?
2. What combination of predictor or independent variables yields the highest prediction accuracy in predicting students' STEM performance?
3. Which of the following data mining techniques - Random Forest, PART, J48 and Naive Bayes provide the best prediction accuracy in predicting students' STEM performance?
4. How accurate will predictions be if selected attributes are used in predictive models of predicting students' STEM performance and how well it predicts academically-at-risk students?

The above questions will be the leading goal in which the research will be design and studied afterwards.

1.5 Significance of The Research

On the basis of the study's findings, the information can be utilized in a variety of methods to alleviate or reduce educational issues in Malaysia. The high influential attributes identified in predicting students' academic performance, such as district, type of school, dual language program (DLP) status, religion, dorm, nationality, guardian's occupation, guardian's salary group, and mid-year exam results, can be used to develop personalized learning plans for students. By understanding these factors, schools are able to tailor educational strategies and interventions to the unique requirements of each student, thereby enhancing their learning outcomes (Hasselbring & Glaser, 2000).

The study's classification algorithms with an average prediction accuracy of over 70% can aid in the early identification of students at academic risk. This information can be used by educational institutions and policymakers to implement targeted intervention programs for students who are more likely to encounter academic difficulties. These interventions may consist of additional tutoring, mentoring, or support services in order to improve students' academic performance and prevent prospective withdrawals (Herman & Yeh, 1980).

By understanding the most influential factors on student performance, institutions and the educational ministry can allocate resources more efficiently. Schools in districts with superior academic performance can share their best practices and resources with schools confronting difficulties, resulting in more equitable educational opportunities and improved student outcomes overall (Hanushek & Rivkin, 2006).

The findings can also be used to inform curriculum development and enhancement strategies. Schools can modify their curricula to address the unique requirements and obstacles of students in various settings. For instance, if the type of school or language program has a significant impact on student performance, schools can implement language support or custom-tailored programs to accommodate students from a variety of backgrounds (Darling-Hammond & Lieberman, 2012).

The study's findings can also be used to foster collaboration between parents, teachers, and the educational ministry. Parents can be made aware of the factors that affect their children's academic performance, allowing them to provide greater support at home (Epstein & Sheldon, 2002). Teachers can use this information to develop more effective teaching strategies by recognizing the diverse requirements of their students and adapting instructional approaches accordingly.

The methodological approach of the study can serve as a foundation for the educational ministry to develop evidence-based policies aimed at enhancing academic performance and student achievement. This data-driven approach can be used by the ministry to develop policies that address the unique challenges encountered by schools and students, ultimately leading to improved educational outcomes (Cohen, 2015).

There are three significant reasons for this study. First, this study will provide a methodological approach that will inform the ministry of education on how to predict students' performance using data available in the Education Data Repository. Fortunately, most of the input characteristics or background data of secondary school students are stored online in secured database of the ministry of education. However, most of the time, while they may have access, they may not necessarily have an approach to follow to make meaning of this data. Most of the systems and applications in the Education Data Repository environment are used at school level only for compliance and monitoring purposes. Compiling individual student information allows head teachers to request data from each teacher responsible for inserting data into each Education Data Repository system. In order to form a representation of the student's background, achievement, and attendance, the head teacher would then collect all data manually. By providing a methodological approach to analyze the students' data, this study will help provide the ministry of education with a procedural process for predicting student performance and analyzing student outcomes.

Benefits to the Ministry of Education from the study's methodology for analyzing student data are considerable. It paves the way for data-driven decision making, which in turn exposes patterns, trends, and trouble spots in the educational system. This method makes it easier to allocate funds efficiently, providing more targeted assistance to children most in need and enhancing educational results for everyone. It also enables continuous monitoring and assessment for progress tracking and improvement, provides early intervention measures to reduce dropouts, and informs evidence-based educational policy. To provide quality assurance and improve students' entire educational experience, a data-driven strategy promotes improved communication with stakeholders and enables long-term planning to successfully tackle future difficulties.

This is aligned with the purpose of the Education Data Repository. As mentioned in the report produce by UNICEF (2020):

> The data derived from the repository are used to inform decision-making and policy. For example, data are analyzed and used to inform parliamentary debates on education policy. When questions are raised in parliament concerning education, such as shortages of teachers in specific areas or subjects, the Minister and officials provide answers based on information that is accessible on the KPM Dashboard. Analyzed data are also a major source of statistical data within the Ministry to make decisions on resourcing to districts and schools, as well as to guide policy and policy implementation. (UNICEF, 2020, p.57)

Second, this study was significant because the results of this study can help inform the ministry of education, teachers and parents with valuable insight on key factors that can be used to help improve academic performance as well as students’ success. These insights can be used to help shape future curriculum policy and allow procedures to be put in place to help promote these factors among students prior to them taking the subjects of interest. The findings of this research have important implications for the success of children and the quality of education provided by the Ministry of Education, schools, and parents. These findings may guide curricular decisions in the future and pave the way for specialized interventions to strengthen foundational skills among students before they ever encounter the necessary coursework. Teachers, parents, and other members of the education community may use this information to work together to improve the learning environment in which students thrive, paving the way for greater academic success in the long run. For this reason, four data mining algorithms were used in this analysis to model the performance of students in related subjects which will be beneficial to educators and students in a lot of aspects.

Third, it is impossible to generalize about students' success in STEM fields without first having access to the information that was gleaned through the study's examination of data mining methods. The purpose of this research is to discover the data mining algorithm that is the most reliable and useful for the early diagnosis of academically at-risk students. This may be accomplished by evaluating the performance of several data mining algorithms using precise accuracy measures such as Precision, Recall, F-measure, and ROC area. As soon as the best model has been selected, it will be possible to use it across a larger student population. This will enable the Ministry of Education to forecast and locate students who are at risk in a variety of schools and geographies.

Making broad generalizations about the performance of students in STEM fields offers various advantages, one of which is the capacity to identify students who are academically at risk early. To begin, it offers a more thorough knowledge of the elements that contribute to the academic troubles of kids. This helps in identifying similar patterns and obstacles that are confronted by students who are at risk in a variety of different circumstances. This knowledge may be put to use in the creation of focused intervention methods and support systems with the purpose of more effectively addressing these difficulties.

Furthermore, if the Ministry of Education has a dependable model for the early prediction of students who will drop out of school, it will be able to proactively take preventive measures to lower the percentage of students who do so. The ministry is able to enhance retention rates and overall academic performance by assisting students who are having difficulty by providing them with timely help and resources.

In addition, the results of the research may provide a representative picture of the general performance of the student population in STEM courses since the research was conducted by analyzing a big and varied dataset using the most effective data mining technique. The Ministry of Education may use this information to make more informed choices on the development of curricula, the allocation of resources, and the implementation of policy in order to enhance the results of STEM education on a wider scale.

1.6 Scope and Limitations of Research

The present study was limited to datasets gathered over from Form 4 students whose taking STEM subjects all over Malaysia. The initial plan of the study is to get the data for all Form 4 students in Malaysia. Unfortunately, because of the confidentiality of the data, the researcher cannot get the data for all Form 4 students from EPRD (Education Planning and Research Division). Alternatively, the researcher gathered the data by contacting all JPN (State Education Department) in Malaysia but again with the confidentiality issue of the data the researcher can only get from the state of Selangor. The outcomes will be much more promising and interesting if datasets from Form 5 can also be gathered. However, it is so difficult to get datasets of SPM results as it is classified. Then, the students' data collected cannot be generalized to all secondary students in Malaysia. This is because the coverage for APDM also does not include schools under the jurisdiction of other governmental agencies. For private schools, the Private Education Information System (SMIPS) is used to collect aggregated data on students (UNICEF, 2020). Other challenges in predicting students' academic performance are selecting the right factors and relevant attributes with a correct prediction method. Student academic performance is affected by numerous factors. The scope of the research is limited to the investigation of the effects of a student's background and their respective results in final examination. Nevertheless, the statistical data mining techniques used in this research can be applicable to other schools or even universities in analyzing their respective student performance data not only in STEM subjects but any subjects or courses and in the field of higher education institutional research.

1.7 Definition of Terms

A variety of words were used extensively in this thesis, and for clarity's sake, meanings of such words are given here.

Secondary schools. Individuals in Malaysia begin secondary school at the age of thirteen. All Malaysian children are required to attend secondary school. Secondary school education typically lasts five years. Lower Secondary (Forms 1 to 3) and Upper Secondary (Forms 4 to 5) are the two levels of secondary education (Ministry of Education Malaysia). In this study, only form 4 students in the category of Upper Secondary will be involved in the analysis.

STEM. STEM is a curriculum based on the idea of educating students in four specific disciplines — science, technology, engineering and mathematics — in an interdisciplinary and applied approach (Lee et al., 2019). In this study, only four main subjects in Science stream will be included in representing STEM performance which are Additional Mathematics, Physics, Chemistry and Biology.

Modeling. According to Witten and Frank (2005), modeling is the step in the data mining process where the patterns identified in the data are used to construct a predictive model, which can then be applied to new data to make predictions or classify new observations. In this study, a set of models is generated utilizing various statistical algorithms or data mining techniques, often known as ensemble models.

At risk students. At risk students are students who have a higher probability of getting lower grades in STEM subjects. The higher education literature defines "at-risk" as a term with origins in K-12 education meaning students that "are poorly equipped to perform up to academic standards" (Quinnan, 1997, p.31). In this study, at risk students are students who obtained examination marks of that below 40 or getting the grade G.

Data Mining. Data mining is defined as the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data ((Sumathi & Sivanandam, 2006). In this study, different data mining techniques (Decision Tree (DT), Artificial Neural Networks (ANN), and Naive Bayes (NB)) will be explored and compared to discover all the hidden knowledge within all data obtained from student data repository.

Educational Data Mining (EDM). Educational Data Mining (EDM) is an area that utilizes the different forms of educational data in statistics, machine learning and data mining (DM) algorithms (Romero & Ventura, 2010). In this study, one of the applications of EDM will be used which is predicting student performance to investigate and understand students and the settings in which they learn in educational settings of the unique types of data.

Academic performance. Academic performance is defined as the observable results and accomplishments that students demonstrate when they are enrolled in educational institutions. It involves both cognitive and non-cognitive areas, such as the acquisition of knowledge, ability in critical thinking and problem-solving, the retention of information, academic engagement, and overall success in academic undertakings. It is standard practice to evaluate students' academic achievement using a variety of different types of assessments, including tests, quizzes, projects, assignments, and exams. (Kuh et al., 2006). In this study, academic performance refers to mid year and final year examination of Form 4 students in terms of their grades.

Variable. Variable is defined as the characteristic or attribute of a student. A variable is a characteristic or attribute of the data that is being analysed. In data mining, variables are used to create models and identify patterns in the data (Witten and Frank, 2005). In this study, the variables involved are District, DLP Status, Sex, Race, Religion, Orphan, Dormitory, OKU, Nationality of Guardian 1, Job of Guardian 1, Income of Guardian 1, No. of Dependents, Nationality of Guardian 2, Combined Income, Salary Group, Mid-Year Attendance, Final Year Attendance, Mid-Year Chemistry, Mid-Year Physics, Mid-Year Biology, Mid-Year Add Maths, Type of School, Final Year Chemistry, Final Year Physics, Final Year Biology, and Final Year Add Maths.

Input factors. Astin (1993) defined inputs as characteristics of the student at the time of entry to the school. In this study, input factors refers to students' demographic and students' academic record in their corresponding schools.

Environment factors. Environment factors are defined as various programs, policies, faculty, peers, and educational experiences to which the student is exposed (Astin, 1993). In this study, environment factors only refer to one variable which is the type of school the students are attending. It can be regular, boarding or religious school.

1.8 Chapter Summary

Chapter 1 starts with the background of the study that will be conducted. The study begins with introducing the Fourth Industrial Revolution (4IR) which mainly centered upon advanced technology where data analytics become very important. Then, also related to 4IR, is the significant of thriving in STEM subjects so that students will survive in the upcoming complex world. Furthermore, schools are heading towards data driven organization where data related to students are stored online. There is an Education Data Repository developed by the Ministry of Education to store the database of students in different kind of systems and applications. This database will be very beneficial to inform not only the ministry, but also schools and parents about students. The most highly anticipated element of this database is of course the achievement or performance of students. This leads to the discussion of data mining where one of the aims in data mining is prediction.

Next, chapter 1 also discuss the problem statement that arises from the background of the study and then the research approach on how to tackle the problem stated. This chapter continues to introduce the context of the research as well as the objectives and research questions to address the objectives. Ultimately, the research significantly needs to be conducted to solve the problem and the important aim is to predict students' academic performance in Malaysian secondary schools.

The rest of the paper is organized as follows. Chapter 2 will describe in details existing educational data science literature that is crucial in predicting students' performance. Chapter 3 describes the dataset and discusses the methodology. Chapter 4 will go through the first phase of DDR which involve three stages of CRISP-DM conceptual model, mainly business understanding, data understanding, and data preparation. Chapter 5 will cover phase 2 of DDR (development of the model) and the only one stage of CRISP-DM conceptual model which is modeling. Next, Chapter 6 of this paper will go through two stages of CRISP-DM which are evaluation and deployment and the last phase of DDR (evaluation of the model).

REFERENCES

Abdulazeez, Y., & Abdulwahab, L. (2019). Application of classification models to predict students' academic performance using classifiers ensemble and synthetic minority over sampling techniques. *Bayero Journal of Pure and Applied Sciences*, *11*(2), 142. https://doi.org/10.4314/bajopas.v11i2.17

Abu Saa, A. (2016). Educational Data Mining & Students' Performance Prediction. *IJACSA) International Journal of Advanced Computer Science and Applications*, *7*(5), 212–220. www.ijacsa.thesai.org

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. In *Technology, Knowledge and Learning* (Vol. 24, Issue 4). Springer Netherlands. https://doi.org/10.1007/s10758-019-09408-7

Akar, S. G. M., & Altun, A. (2020). Individual Differences in Learning Computer Programming: A Social Cognitive Approach. *Contemporary Educational Technology*, *8*(3), 195–213. https://doi.org/10.30935/cedtech/6196

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, *16*(1). https://doi.org/10.1186/s41239-019-0172-z

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, *6*(7), 528–533. https://doi.org/10.7763/ijiet.2016.v6.745

Ali, G., Jaaffar, A. R., & Ali, J. (2021). STEM Education in Malaysia: Fulfilling SMEs' Expectation. In *Modeling Economic Growth in Contemporary Malaysia* (pp. 43–57). Emerald Group Publishing Ltd. https://doi.org/10.1108/978-1-80043-806-420211005

Al-Rofo, M. A. (2010). The Dimensions That Affect the Students' Low Accumulative Average in Tafila Technical University. *Journal of Social Sciences*, *22*(1), 53–59. https://doi.org/10.1080/09718923.2010.11892784

Alturki, S., Alturki, N., & Stuckenschmidt, H. (2021). Using Educational Data Mining To Predict Students' Academic Performance For Applying Early Interventions. *Journal of Information Technology Education: Innovations in Practice*, *20*, 121–137. https://doi.org/10.28945/4835

Aluko, R. O., Daniel, E. I., Shamsideen Oshodi, O., Aigbavboa, C. O., & Abisuga, A. O. (2018). Towards reliable prediction of academic performance of architecture students using data mining techniques. *Journal of Engineering, Design and Technology*, *16*(3), 385–397. https://doi.org/10.1108/JEDT-08-2017-0081

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-0177-7

Amrieh, A., Hamtini, T., & Aljarah, I. (2016). *Mining Educational Data to Predict Student's academic Performance using Ensemble Methods*. *9*(8), 119–136.

Asiah, M., Nik Zulkarnaen, K., Safaai, D., Nik Nurul Hafzan, M. Y., Mohd Saberi, M., & Siti Syuhaida, S. (2019). A Review on Predictive Modeling Technique for Student Academic Performance Monitoring. *MATEC Web of Conferences*, *255*, 03004. https://doi.org/10.1051/matecconf/201925503004

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting Student Academic Performance at Degree Level: A Case Study. *International Journal of Intelligent Systems and Applications*, *7*(1), 49–61. https://doi.org/10.5815/ijisa.2015.01.05

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. Journal of College Student Personnel, 25(4), 297-308.

Astin, A. W., & Antonio, A. L. (2012). Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education. Rowman & Littlefield Publishers.

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. *Procedia Computer Science*, *82*, 80–89. https://doi.org/10.1016/j.procs.2016.04.012

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Claudia Goldin and Lawrence F. Katz (2007), The Race between Education and Technology, NBER Working Paper No. 12984. *Mis Quartely*, *40*(4), 807–818.

Bahrum, S., Wahid, N., & Ibrahim, N. (2017). Integration of STEM Education in Malaysia and Why to STEAM. *International Journal of Academic Research in Business and Social Sciences*, *7*(6), 645–654. https://doi.org/10.6007/ijarbss/v7-i6/3027

Bahovec, V., Erjavec, N., & Cizmesija, M. (2013). Logistical analysis of the impact of contributing factors on the success of students of higher education in quantitative courses. Croatian Operational Research Review, 4, 284-294.

Bandura, A. (1986). Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall.

Barbosa Manhães, L. M., da Cruz, S. M. S., & Zimbrão, G. (2015). *Towards automatic prediction of student performance in STEM undergraduate degree programs*. 247–253. https://doi.org/10.1145/2695664.2695918

Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive models of student college commitment decisions using machine learning. *Data*, *4*(2), 1–18. https://doi.org/10.3390/data4020065

Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: An introductory overview. In Proceedings of the 3rd International Conference on Research Challenges in Information Science (RCIS 2009) (pp. 439-446). IEEE. doi:10.1109/RCIS.2009.5089308

Begum, S., & Padmannavar, S. S. (2022). Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction. *International Journal of Intelligent Engineering and Systems*, *15*(2), 316–328. https://doi.org/10.22266/ijies2022.0430.29

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers and Education*, *158*(September 2019), 103999. https://doi.org/10.1016/j.compedu.2020.103999

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology, 66(11), 2215-2222. doi:10.1002/asi.23329

Boote, D. N., & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature review in research preparation. Educational Researcher, 34(6), 3-15. doi:10.3102/0013189X034006003

Božić, R. (2022). Predicting Study Programme Selection with Data Mining Classification Technique. *Ijeec - International Journal of Electrical Engineering and Computing*, *5*(2), 94–101. https://doi.org/10.7251/ijeec2102094b

Brisk, M. E. (Boston C. (1999). Quality BIlingual Education: Defining Success. *Symposium on Language Policy*, *1*.

Bronfenbrenner, U. (1979). The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press.

Brookshire, R. G., & Palocsay, S. W. (2005). Factors contributing to the success of undergraduate business students in management science courses. Decision Sciences Journal of Innovative Education, 3(1), 99-108. doi:10.1111/j.1540- 4609.2005.00054.x

Buck Institute for Education. (2018). Project Based Learning Research Review. https://www.pblworks.org/why-pbl/research

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*, *66*, 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Chango, W., Cerezo, R., & Romero, C. (2021). Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Computers and Electrical Engineering*, *89*(October 2020). https://doi.org/10.1016/j.compeleceng.2020.106908

Chaudhari, K. P., Sharma, R. A., Jha, S. S., & Bari, R. J. (2017). Student Performance Prediction System using Data Mining Approach. *Ijarcce*, *6*(3), 833–839. https://doi.org/10.17148/ijarcce.2017.63195

Chen, X., & Ran, W. (2018). The Integration of Technology into Teaching: A Case Study of Two Secondary Schools in China. International Journal of Emerging Technologies in Learning, 13(10), 124-136.

Cohen, D. K. (2015). Teaching and its predicaments. Harvard University Press.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Creswell, J. W. (2012). Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research. (Vols. 4, revised). Pearson, 2012.

Cronin, J., McKenzie, G., & Stirling, A. (2018). Towards a new indicator of research article impact: RCR—relative citation ratio. Journal of Informetrics, 12(2), 533-542. doi:10.1016/j.joi.2018.03.004

Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, *47*(xxxx), 5260–5267. https://doi.org/10.1016/j.matpr.2021.05.646

Darling-Hammond, L., & Lieberman, A. (2012). Teacher evaluation in the United States: A comparative perspective. Learning Policy Institute.

Daud, A., Lytras, M. D., Aljohani, N. R., Abbas, F., Abbasi, R. A., & Alowibdi, J. S. (2019). Predicting student performance using advanced learning analytics. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 415–421. https://doi.org/10.1145/3041021.3054164

Davis, S., Connolly, A., & Linfield, E. (2009) Lecture capture: Making the most of faceto-
face learning, Engineering Education, 4(2), 4-13

Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research*, *281*(3), 575–587. https://doi.org/10.1016/j.ejor.2019.03.037

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Du, X., Yang, J., Hung, J. L., & Shelton, B. (2020). Educational data mining: a systematic review of research and emerging trends. *Information Discovery and Delivery*, *48*(4), 225–236. https://doi.org/10.1108/IDD-09-2019-0070

Dweck, C. S. (2006). Mindset: The new psychology of success. Random House.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. Annual Review of Psychology, 53(1), 109-132.

Ellis, T. J., & Levy, Y. (2008). Framework of problem-based research: A guide for novice researchers on the development of a research-worthy problem. *Informing Science*, *11*, 17–33. http://www.inform.nu/Articles/Vol11/ISJv11p017-033Ellis486.pdf

Epstein, J. L., & Sheldon, S. B. (2002). Present and accounted for: Improving student attendance through family and community involvement. The Journal of Educational Research, 95(5), 308-318.

European Commission. (2012). Science, Technology, Engineering and Mathematics (STEM) in the European Education Area. Publications Office of the European Union.

Ezz, M., & Elshenawy, A. (2020). Adaptive recommendation system using machine learning algorithms for predicting student's best academic program. *Education and Information Technologies*, *25*(4), 2733–2746. https://doi.org/10.1007/s10639-019-10049-7

F. Gontzis, A., Kotsiantis, S., T. Panagiotakopoulos, C., & Verykios, V. S. (2022). A predictive analytics framework as a countermeasure for attrition of students. *Interactive Learning Environments*, *30*(3), 568–582. https://doi.org/10.1080/10494820.2019.1674884

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Data Mining and Knowledge Discovery in Databases. *Communications of the ACM*, *39*(11), 24–26. https://doi.org/10.1145/240455.240463

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*(August 2017), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Fullan, M., & Langworthy, M. (2014). A Rich Seam: How New Pedagogies Find Deep Learning. London: Pearson.

Galvan, P. (2016). Educational Evaluation and Prediction of School Performance through Data Miningand Genetic Algorithms. *International Journal of Advanced Engineering Research and Science*, *3*(10), 215–220. https://doi.org/10.22161/ijaers/3.10.34

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38(4), 915-945.

Gao, L., Zhao, Z., Li, C., Zhao, J., & Zeng, Q. (2022). Deep cognitive diagnosis model for predicting students' performance. *Future Generation Computer Systems*, *126*, 252–262. https://doi.org/10.1016/j.future.2021.08.019

Gardiner, E. J., & Gillet, V. J. (2015). Perspectives on Knowledge Discovery Algorithms Recently Introduced in Chemoinformatics: Rough Set Theory, Association Rule Mining, Emerging Patterns, and Formal Concept Analysis. *Journal of Chemical Information and Modeling*, *55*(9), 1781–1803. https://doi.org/10.1021/acs.jcim.5b00198

Garg, R. (2018). *Predicting Student Performance Of Different Regions Of Punjab Using Classification Techniques*. *9*(1), 236–240.

Gough, D., Oliver, S., & Thomas, J. (2017). An introduction to systematic reviews (2nd ed.). SAGE Publications Ltd.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information and Libraries Journal, 26(2), 91-108. doi:10.1111/j.1471-1842.2009.00848.x

Guabassi, I. El, Bousalem, Z., Marah, R., & Qazdar, A. (2021). Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance. *International Journal of Online and Biomedical Engineering*, *17*(2), 90–105. https://doi.org/10.3991/ijoe.v17i02.20025

Halim, L.; Meerah, T. S. M. (2016). Science education research and practice in Asia: Challenges and opportunities. *Science Education Research and Practice in Asia: Challenges and Opportunities*, 1–578. https://doi.org/10.1007/978-981-10-0847-4

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. In E. A. Hanushek & F. Welch (Eds.), Handbook of the economics of education (Vol. 2, pp. 1051-1078). North-Holland.

Harvard Graduate School of Education. (2015). Making in America: From Innovation to Market. http://www.gse.harvard.edu/news/15/01/making-america-innovation-market

Hasselbring, T. S., & Glaser, R. (2000). Using Student Data to Assess the Impact of Technology on Teaching and Learning. Journal of Research on Technology in Education, 33(1), 5-15.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, *161*, 134–146. https://doi.org/10.1016/j.knosys.2018.07.042

Herman, R., & Yeh, S. S. (1980). Synthesis of research on student factors related to test performance. Educational Testing Service.

Herrmannova, D., & Knoth, P. (2016). Semantometrics: Fulltext-based measures for analyzing research publications. Scientometrics, 109(2), 1291-1322. doi:10.1007/s11192-016-2077-6

Hirji, K. K. (2001). Exploring Data Mining Implementation. *Mining of Massive Datasets*, *2*, 87–93. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

Hirokawa, S. (2018). Key attribute for predicting student academic performance. *ACM International Conference Proceeding Series*, *March*, 308–313. https://doi.org/10.1145/3290511.3290576

Huang, R., Spector, J. M., & Yang, J. (2019). Educational Technology a Primer for the 21st Century. In *Journal of Educational Television* (Vol. 1, Issue 2). https://doi.org/10.1080/1358165750010212

Hung, H. C., Liu, I. F., Liang, C. T., & Su, Y. S. (2020). Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry*, *12*(2), 1–14. https://doi.org/10.3390/sym12020213

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

International Association for the Evaluation of Educational Achievement (IEA). (2021). TIMSS. Retrieved from https://timssandpirls.bc.edu/

Imran, M., & Afshari, M. (2013). Demography Factors and Students? Academic Performance in Secondary Schools. *Malaysian Online Journal of Educational Management*, *1*(2), 1–9.

Injadat, M. N., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. *Applied Intelligence*, *50*(12), 4506–4528. https://doi.org/10.1007/s10489-020-01776-3

Jayaprakash, S., & Jaiganesh, V. (2019). Predicting academic performance of tertiary students using classification algorithm. *International Journal of Recent Technology and Engineering*, *8*(2), 6558–6561. https://doi.org/10.35940/ijrte.B2716.078219

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences (Switzerland)*, *11*(7). https://doi.org/10.3390/app11073130\

Kaighobadi, M., & Allen, M. T. (2008). Investigating academic success factors for undergraduate business students. Decision Sciences Journal of Innovative Education, 6(2), 427-436. doi:10.1111/j.1540-4609.2008.00184.x

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(1). https://doi.org/10.1186/s40594-016-0046-z

Keržič, D., Aristovnik, A., Tomaževič, N., & Umek, L. (2019). Assessing the impact of students' activities in e-courses on learning outcomes: a data mining approach. *Interactive Technology and Smart Education*, *16*(2), 117–129. https://doi.org/10.1108/ITSE-09-2018-0069

Khanna, L., Singh, S. N., & Alam, M. (2017). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. *India International Conference on Information Processing, IICIP 2016 - Proceedings*. https://doi.org/10.1109/IICIP.2016.7975354

Kishan Das Menon, H., & Janardhan, V. (2020). Machine learning approaches in education. *Materials Today: Proceedings*, *43*(xxxx), 3470–3480. https://doi.org/10.1016/j.matpr.2020.09.566

Kostopoulos, G., Kotsiantis, S., Fazakis, N., Koutsonikos, G., & Pierrakeas, C. (2019). A Semi-Supervised Regression Algorithm for Grade Prediction of Students in Distance Learning Courses. *International Journal on Artificial Intelligence Tools*, *28*(4), 1–19. https://doi.org/10.1142/S0218213019400013

Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., & Gravvanis, G. A. (2018). Forecasting students' success in an open university.

International Journal of Learning Technology, *13*(1), 26–43. https://doi.org/10.1504/IJLT.2018.091630

Kuh, G. D., Kinzie, J., & Buckley, J. A. (2006). What Matters to Student Success : A Review of the Literature Spearheading a Dialog on Student Success. *Commissioned Report for the National Symposium on Postsecondary Student Success Spearheading a Dialog on Student Success*, *18*(July), 156. http://cpe.ky.gov/NR/rdonlyres/AFA304F0-C125-40C2-96E5-7A8C98915797/0/WhatMatterstoStudentSuccessAReviewoftheLiterature.pdf

Lange, C. (2019). Academic libraries and research data services: Current practices and plans for the future. College & Research Libraries, 80(6), 787-813. doi:10.5860/crl.80.6.787

Lee, M. H., Chai, C. S., & Hong, H. Y. (2019). STEM Education in Asia Pacific: Challenges and Development. In *Asia-Pacific Education Researcher* (Vol. 28, Issue 1). Springer. https://doi.org/10.1007/s40299-018-0424-z

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning Analytics: Fundaments, Applications, and Trends. *Learning Analytics: Fundaments, Applications, and Trends, Studies in Systems, Decision and Control*, *94*(February), 1–23. https://doi.org/10.1007/978-3-319-52977-6

Levy, Y., & Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *IInforing Science: The International Journal of an Emerging Transdiscipline*, *9*, 181–212. http://www.inform.nu/Articles/Vol9/V9p181-212Levy99.pdf

Li, H., Ramanathan, S., Nuthalapati, V., & Lingam, R. (2017). China and India's Race for STEM Talent: A View from Graduate International Students. Journal of International Students, 7(3), 701-725.

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. *Journal of Educational Computing Research*, *57*(2), 448–470. https://doi.org/10.1177/0735633117752614

Lowis, M., & Castley, A. (2008). Factors affecting student progression and achievement: Prediction and intervention. A two-year study. *Innovations in Education and Teaching International*, *45*(4), 333–343. https://doi.org/10.1080/14703290802377232

Machi, L. A., & McEvoy, B. T. (2016). The literature review: Six steps to success (3rd ed.). SAGE Publications, Inc.

Maheswari, K., Priya, A., Balamurugan, A., & Ramkumar, S. (2021). Analyzing student performance factors using KNN algorithm. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2020.12.1024

Mai, T. T., Bezbradica, M., & Crane, M. (2022). Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data. *Future Generation Computer Systems*, *127*(September 2021), 42–55. https://doi.org/10.1016/j.future.2021.08.026

Ministry of Education Malaysia. (2013). Malaysia Education Blueprint 2013-2025: Preschool to Post-Secondary Education. Putrajaya: Ministry of Education Malaysia.

Ministry of Education Malaysia. (2019). Quick Facts 2019. Putrajaya: Ministry of Education Malaysia.

Malini, J., & Kalpana, Y. (2021). Investigation of factors affecting student performance evaluation using education materials data mining technique. *Materials Today: Proceedings*, *47*(xxxx), 6105–6110. https://doi.org/10.1016/j.matpr.2021.05.026

Mannila, H. (2000). Theoretical frameworks for data mining. *ACM SIGKDD Explorations Newsletter*, *1*(2), 30–32. https://doi.org/10.1145/846183.846191

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2013). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. US Department of Education.

Merchán Rubiano, S., Beltrán Gómez, A., & Duarte García, J. (2017). Engineering Students´ Academic Performance Prediction using ICFES Test Scores and Demo-graphic Data. *Ingeniería Solidaria*, *13*(21), 53–61. https://doi.org/10.16925/in.v13i21.1729

Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, *115*, 36–51. https://doi.org/10.1016/j.dss.2018.09.001

Mokhtar, S., Alshboul, J. A. Q., & Shahin, G. O. A. (2019). Towards Data-driven Education with Learning Analytics for Educator 4.0. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1339/1/012079

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355–371. https://doi.org/10.1023/A:1018755201899

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. *IEEE Access*, *9*, 140731–140746. https://doi.org/10.1109/ACCESS.2021.3119596

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press.

Nurhana Roslan, Jastini Mohd Jamil, I. N. M. S. (2021). Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 2326–2334. https://doi.org/10.17762/turcomat.v12i3.1219

OECD. (2019). PISA 2018 Results (Volume I): What Students Know and Can Do. https://doi.org/10.1787/5f07c754-en

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management and Data Systems*, *116*(8), 1678–1699. https://doi.org/10.1108/IMDS-09-2015-0363

Pacific, U. E. A. &. (2020). *INFORMATION SYSTEMS (EMIS)* (Issue July). https://www.unicef.org/eap/reports/review-education-management-information-systems-emis-track-individual-student-data

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile. *Entropy*, *23*(4), 1–23. https://doi.org/10.3390/e23040485

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2021). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2021.07.382

Perner, P. (2015). Machine Learning and Data Mining in Pattern Recognition 11th International Conference, MLDM 2015 Hamburg, Germany, July 20-21, 2015 Proceedings. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9166*, 403–414. https://doi.org/10.1007/978-3-319-21024-7

Qatar Foundation. (2021). Education City: Unlocking Human Potential. Qatar Foundation.

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, *24*(6), 3577–3589. https://doi.org/10.1007/s10639-019-09946-8

Ramaswami, G., Susnjak, T., Mathrani, A., Lim, J., & Garcia, P. (2019). Using educational data mining techniques to increase the prediction accuracy of student academic performance. *Information and Learning Science*, *120*(7–8), 451–467. https://doi.org/10.1108/ILS-03-2019-0017

Reeves, T. C., Herrington, J., & Oliver, R. (2005). Design research: A socially responsible approach to instructional technology research in higher education. Journal of Computing in Higher Education, 16(2), 97-116. doi:10.1007/BF02961480

Richey, R. C., & Klein, J. D. (2007). *Design and Development Research*. Routledge.

Riestra-González, M., Paule-Ruíz, M. del P., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers and Education*, *163*(July 2020). https://doi.org/10.1016/j.compedu.2020.104108

Rocco, T. S., & Plakhotnik, M. S. (2009). Literature reviews, conceptual frameworks, and theoretical frameworks: Terms, functions, and distinctions. Human Resource Development Review, 8(1), 120-130. doi:10.1177/1534484308326284

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. 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*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Roslan, M. H. bin, & Chen, C. J. (2022). Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). *International Journal of Emerging Technologies in Learning*, *17*(5), 147–179. https://doi.org/10.3991/ijet.v17i05.27685

Sahlberg, P. (2011). Finnish Lessons: What can the world learn from educational change in Finland? Teachers College Press.

Samundeeswari, K., & Dr.K.Srinivasan. (2017). Data Mining Techniques In Agriculture Prediction Of Soil Fertility. *International Journal of Scientific & Engineering Research*, *8*(4), 45–51. http://www.ijser.org

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. *Internet and Higher Education*, *37*, 76–89. https://doi.org/10.1016/j.iheduc.2018.02.002

Sergis, S., Sampson, D. G., Rodríguez-Triana, M. J., Gillet, D., Pelliccione, L., & de Jong, T. (2019). Using educational data from teaching and learning to inform teachers’ reflective educational design in inquiry-based STEM education. *Computers in Human Behavior*, *92*, 724–738. https://doi.org/10.1016/j.chb.2017.12.014

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015a). A Review on Predicting Student’s Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015b). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Sharma, P. (2019). Digital Revolution of Education 4.0. *International Journal of Engineering and Advanced Technology*, *9*(2), 3558–3564. https://doi.org/10.35940/ijeat.a1293.129219

Sharma, Y. K. (2002). *Fundamental aspects of educational technology*.

Shingari, I., & Kumar, D. (2018). Predicting Student Performance Using Classification Data Mining Techniques. In *International Journal of Computer Sciences and Engineering* (Vol. 6, Issue 7, pp. 43–48). https://doi.org/10.26438/ijcse/v6i7.4348

Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Sumathi, S., & Sivanandam, S. N. (2006). Introduction to Data Mining and its Applications. In *Studies in Computational Intelligence* (Vol. 29). https://doi.org/10.1007/978-3-540-34351-6

Tegegne, A. K., & Alemu, T. A. (2018). Educational data mining for students' academic performance analysis in selected Ethiopian universities. *Information Impact: Journal of Information and Knowledge Management*, *9*(2), 1–15. https://doi.org/10.4314/iijikm.v9i2.1

Terry, N., Macy, A., Clark, R., & Sanders, G. (2015). The impact of lecture capture on student performance in business courses. Journal of College Teaching & Learning, 12(1), 65-74.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89-125.

Tytler, R., Osborne, J., Williams, G., Tytler, K., & Cripps Clark, J. (2008). Opening up pathways: Engagement in STEM across the primary-secondary school transition. Australian Government Department of Education, Employment and Workplace Relations.

UNESCO. (2017). Cracking the code: Girls' and women's education in science, technology, engineering, and mathematics (STEM). https://unesdoc.unesco.org/ark:/48223/pf0000245656

Upadhyay, N. (2014). EDM-65: A Survey on the Classification Techniques In Educational Data Mining. *India - WEKA*, *3*(11), 725–728.

Vanthienen, J., & De Witte, K. (Eds.). (2017). D. analytics applications in education. C. Press. (n.d.). *Data Analytics Applications in Education* (2018th ed.). Taylor & Francis Group.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189

World Economic Forum. (2016). The Fourth Industrial Revolution. Geneva: World Economic Forum.

Wouters, P., & Costas, R. (2012). Users, narcissism and control—tracking the impact of scholarly publications in the 21st century. Utrecht University. Retrieved from https://dspace.library.uu.nl/bitstream/handle/1874/257870/WoutersCostasUsersNarcissismControl.pdf?sequence=1

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*(January), 166–173. https://doi.org/10.1016/j.chb.2019.04.015

Yousef, D. A. (2011). Academic performance of business students in quantitative courses: A study in the faculty of business and economics at the UAE university: Academic performance of business students in quantitative courses. Decision Sciences Journal of Innovative Education, 9(2), 255-267. doi:10.1111/j.1540- 4609.2011.00305.x

Yusuf, A., & John, A. (2019). Classifiers ensemble and synthetic minority oversampling techniques for academic performance prediction. *International Journal of Informatics and Communication Technology (IJ-ICT)*, *8*(3), 122. https://doi.org/10.11591/ijict.v8i3.pp122-127

Zairul, M. (2020). A Thematic Review on Student Centred Learning in the Studio Education. Journal of Critical Reviews, 7 (2), 504 511. https://doi.org/10.31838/jcr.07.02.95

Zulkifli, F., Mohamed, Z., & Azmee, N. A. (2019). Systematic research on predictive models on students' academic performance in higher education. *International Journal of Recent Technology and Engineering*, *8*(2 Special Issue 3), 357–363. https://doi.org/10.35940/ijrte.B1061.0782S319