

# Analyzing Enrolment Patterns: Stacked Ensemble Statistical Learning-Based Approach to Educational Decision Making

**Dr. Zun Liang Chuan** (✉ [chuanzl@umpisa.edu.my](mailto:chuanzl@umpisa.edu.my))

Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuhr Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang Darul Makmur, Malaysia

**Mr. Chong Teak Wei**

Ever AI Holdings Sdn Bhd, 12, Jalan Anggerik Aranda 31/170C, Kota Kemuning, Shah Alam, Selangor Darul Ehsan, Malaysia

**Mr. Nursultan Japashov**

Educational Theory and Practice Department, University at Albany, New York State University, 1400 Washington Ave, Albany, NY 12222, United States of America

**Mr. Soon Kien Yuan**

Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuhr Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang Darul Makmur, Malaysia

**Mr. Tan Wei Qing**

Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuhr Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang Darul Makmur, Malaysia

**Prof. Dr. Noriszura Ismail**

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

**Assoc. Prof. Dr. Choong-Yeun Liong**

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

**Ms. Tan Ee Hiae**

Sekolah Menengah Kebangsaan Semambu, Depan Qtrs Loji Air Semambu, 25350 Kuantan, Pahang Darul Makmur, Malaysia

---

## Research Article

**Keywords:** Additional Mathematics, Enrolment determinants, Statistical learning-based algorithm, Educational policy, Gender equality

**Posted Date:** December 13th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3723176/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

In Malaysia, Additional Mathematics, equivalent to A-level mathematics, played a vital role in Science, Technology, Engineering, and Mathematics (STEM) education. However, a notable decline in enrolment for the Malaysian Certificate of Education's (SPM's) Additional Mathematics subject has raised concerns about the implications for Malaysia's STEM workforce and its role in sustainable economic growth. The study's primary objectives were to identify the determinants that impacted urban upper-secondary students' enrolment in Additional Mathematics within the Kuantan District, Pahang, Malaysia, and to develop a novel stacked ensemble machine learning algorithm based on these determinants, following the CRISP-DM data science methodology. To pursue these objectives, this study collected and analyzed 389 responses from the first-batch urban upper-secondary students in the Kuantan District who had enrolled in the newly revised Standard Based Curriculum for Secondary Schools (KSSM's) Additional Mathematics syllabus, utilizing a modified research questionnaire and a one-stage cluster sampling technique. The findings revealed that determinants such as education disciplines, ethnicity, gender, mathematics self-efficacy, peer influence, and teacher influence had significantly impacted students' decisions to enroll in Additional Mathematics. Moreover, the introduction of the novel stacked ensemble machine learning algorithm had improved predictive accuracy compared to traditional dichotomous logistic regression algorithms on average, particularly at optimal training-to-test ratios of 70:30, 80:20, and 90:10. These insights were valuable for shaping educational policy and practice, emphasizing the importance of promoting STEM education initiatives and encouraging educators and counselors to empower students to pursue STEM careers while actively promoting gender equality within STEM fields.

## Introduction

The global labor market has a growing demand for Science, Technology, Engineering, and Mathematics (STEM) professionals, including statistical learning-based engineers, artificial intelligence specialists, motion designers, Virtual Reality (VR) professionals, and wind turbine service technicians. According to the United States Department of Labor's employment projections, STEM occupations, which encompass roles in computer and mathematical fields, architecture and engineering, life and physical science, managerial positions, postsecondary teaching, and sales roles requiring scientific or technical knowledge at the postsecondary level, are expected to experience significant growth of 10.8 percent from 2022 to 2032. In contrast, non-STEM occupations are projected to grow by merely 2.3 percent during a similar period (United States Department of Labor, 2023). This data underscores the critical importance of STEM education and careers, both of which rely heavily on a solid foundation in mathematics.

Particularly, improving enrolment in STEM careers for students is crucial for the sustainability of national economic development and growth (Razali et al., 2017; Shahali et al., 2017; Bowden et al., 2018; Halim et al., 2018; Ramli & Awang, 2020; Siregar et al., 2023). In Malaysia, STEM education was formally integrated into the secondary education system in 2020. This transformation involved replacing the Integrated Secondary School Curriculum (KBSM) with the revised Standard Based Curriculum for Secondary Schools (KSSM), in alignment with the Malaysia Education Blueprint 2013–2025 (Malay Mail,

2016). The key differences between KBSM and KSSM curricula can be summarized into five principal pillars: communication, physical and aesthetics, self-directed learning, STEM, and spirituality, attitudes, and values (Dom, 2019).

The KSSM curriculum in Malaysia emphasizes student-centered teaching, problem-solving skills, project-based assignments, regular updates on subjects, and formative assessments (Malay Mail, 2016). However, there has been a significant decline in STEM enrolment among upper-secondary students in Malaysia, dropping from 45.20 percent in 2017 to 40.94 percent in 2022, deviating from the country's targeted science and arts ratio of 60:40 (also known as STEM: Non-STEM ratio in Malaysia Education Blueprint 2013–2025), a policy in place since 1967 (Ramli, & Awang, 2020; Bernama, 2021; The Vibes, 2023; Wen, 2023). This trend is not unique to Malaysia but is a global issue (Halim et al., 2018; Mohtar et al., 2019). Consequently, a majority of literature focuses on exploring determinants that impact secondary school student's interest in STEM education and careers, with a particular emphasis on secondary school students due to their diminishing interest since early schooling, resulting in the reduced pursuit of tertiary STEM education (Shahali et al., 2017; Kamsi et al., 2019).

In Malaysia's education system, Additional Mathematics, equivalent to A-level Mathematics, is regarded as advanced mathematics in upper-secondary education. It serves as an elective subject within STEM education and acts as a typical indirect prerequisite for enrolling in tertiary STEM-related programs at Malaysian public universities. This is due to this subject's coverage of fundamental mathematics knowledge, including algebra, geometry, calculus, trigonometry, and statistics (Ministry of Education Malaysia, 2019).

In recent years, there has been a significant decline in the enrolment of upper-secondary students in Additional Mathematics in Pahang state. According to data from the Pahang State Education Department, enrolment dropped from 7958 students to 5523 students in 2014–2018 (Hiae, 2020). This declining trend also extends to the number of registered candidates for the Malaysian Certificates of Education (SPM) exams, with some students withdrawing from Additional Mathematics before registering for the SPM. Moreover, Pahang State's STEM-related labor force statistics, as reported by the Department of Statistics Malaysia (2023), rank among the bottom five states in Malaysia in 2010–2022 on average. This trend is consistent when considering the number of STEM enrolments, both in Peninsular Malaysia and the East-Coast region, despite Pahang being the largest state in both areas. Notably, Pahang's average STEM-related labor force is merely one-eighth of that in Selangor, the top-ranked state in this regard.

Consequently, the primary objective of this article is to investigate the statistically significant determinants that impacted the upper-secondary enrolment in Additional Mathematics by utilizing appropriate statistical methods. This qualitative research aims to unlock the root causes corresponding to this decline and contribute to achieving the national 60:40 STEM:non-STEM ratio fixed in national education policy. Additionally, this study seeks to develop a robust stacked ensemble statistical learning-based algorithm (SESLA) for predicting Additional Mathematics enrolment based on these determinants

utilizing Cross Industry Standard Process for data mining (CRISP-DM) data science methodology. This study focuses specifically on urban upper-secondary students in the Pahang Kuantan District (Pahang's capital and largest city), who represent the first batch of students enrolled in the KSSM curriculum for Additional Mathematics. By addressing the enrolment decline and understanding its causes, this research aims to unlock insights into improving enrolment and interest in Additional Mathematics among students in Pahang and contribute to the broader national educational goals for prosperous and sustainable national economic growth.

## Literature Reviews

Given the sparse availability of the previous research on the determinants impacting the enrolment of upper-secondary students in Additional Mathematics studies in Malaysia (Authors, 2021), the authors proposed identifying statistically significant determinants individually utilizing a non-parametric chi-squared test. However, this statistical method has limitations as it does not account for adjustment effects among the determinants and potential confounding effects, thus failing to provide comprehensive insights. Therefore, this section shifts its focus to the determinants impacting students' interest in STEM education and careers, drawing inspiration from existing literature.

For instance, Blotnicky et al. (2018) investigated determinants impacting middle school students' enrolment in STEM careers in Atlantic Canada. They employed a logistic regression algorithm to analyze the association between STEM career knowledge, mathematics self-efficacy, grade level, career interests, and engagement in STEM career activities. Their findings indicated that higher STEM career knowledge and mathematics self-efficacy were associated with an increased likelihood of pursuing STEM careers. Moreover, students interested in technical and scientific skills were more inclined towards STEM careers compared to those with preferences for practical, productive, and concrete activities.

In a related study in the United States of America (USA), Bowden et al. (2018) investigated the impact of parental occupation and gender differences on students' mathematics performance utilizing a random-effect panel regression algorithm. Their results demonstrated that students' standardized mathematics test performance was impacted by the STEM occupations of their parents, with gender differences observed.

Similarly, Kaleva et al. (2022) investigated the association between upper-secondary and tertiary students' mathematics choices, university admissions, and gender diversity in STEM career pathways in Finland. Their findings revealed that university admissions significantly impact students' mathematics choices, with more male students opting for advanced mathematics compared to females. This difference was attributed to female students' perceived lack of mathematics self-efficacy, ability, and competence. Males expressed the greatest interest in Information Technology (IT), IT Communication, and Technology, while females showed more interest in Health and Wellbeing, and Education. However, it's worth noting that the study's utilization of the parametric  $t$ -test on ordinal-level Likert scale data was inappropriate, as Likert scale data is qualitative and not suitable for parametric testing.

In Indonesia, Siregar et al. (2023) investigated the impact of gender and parental education on STEM interests utilizing Analysis of Variance (ANOVA) and Likert scale mean scores. However, a limitation of this study was the utilization of ANOVA for analyzing Likert scale scores, which can lead to misleading conclusions due to the ordinal nature of the data. ANOVA also failed to account for adjustments among the determinants and potential confounding effects, raising concerns about the accuracy of the insights derived from the collected data.

In a series of studies conducted in Malaysia, various factors impacting students' STEM education and career interests were investigated. These studies focused on attitude, motivation, parental influence, and the effectiveness of intervention programs. However, some concerns have been raised about the data analysis and interpretation methods utilized in these studies. For instance, Razali et al. (2017) investigated determinants impacting students' STEM career interest among science-stream students in Selangor, emphasizing attitude, motivation, and parental influence. Concerns arose due to the inappropriate utilization of Exploratory Factor Analysis (EFA) in conjunction with Principal Component Analysis (PCA) and arithmetic mean on ordinal-level Likert scale data, which may lead to misinterpretation. Similarly, Shahali et al. (2017) conducted a quasi-experimental study evaluating the effectiveness of the project-based engineering design process in Perak and *Universiti Kebangsaan Malaysia*, Selangor. Their finding showed that the intervention program increased students' interest in STEM education and careers. However, like previous studies, summarized and concluded based on ordinal-level Likert scale data utilizing arithmetic mean, which may lead to misleading interpretations.

Additionally, Halim et al. (2018) investigated STEM self-efficacy's impact on STEM and physics careers across six regions, revealing gender-based variations with male students displaying high self-efficacy in engineering and female students excelling in science disciplines. Boarding school students also exhibited a heightened interest in STEM careers compared to those from public schools. Nonetheless, concerns persist regarding the adequacy of their data analysis and interpretation, echoing issues found in Siregar et al.'s (2023) study. Meanwhile, Kamsi et al. (2019) explored determinants of students' enrolment in STEM education in Perak, utilizing PCA and multiple linear regression (MLR). Their results highlighted the significance of low morale attitude, learning experience, and return on education investment in STEM interest. It's important to note that the utilization of PCA and MLR for a categorical endogenous response raises statistical concerns affecting result interpretability.

In Mohtar et al.'s (2019) study, Structural Equation Modelling (SEM) was employed to investigate determinants impacting interest in physical science and life science careers across six regions. Their finding revealed that STEM self-efficacy positively impacted both physical and life sciences STEM career interests, while perceptions of STEM careers primarily impacted interest in life sciences STEM careers. Nevertheless, concerns were raised about their utilization of inappropriate statistical methods, such as arithmetic mean and normal distribution, with Likert scale data, which is considered qualitative.

Conversely, Ramli & Awang (2020) explored the determinants impacting the implementation of Malaysia's 60:40 STEM Education Policy through semi-structured interviews. They identified determinants at the

students, school, parents, and administrator levels, including determinants like interest, self-confidence, motivation, teacher influence, laboratory infrastructure, parents' perceptions of career opportunities, and administrator qualifications. However, it's important to note that this study presented subjective conclusions without statistical analysis, and relied on data from a small sample of eight respondents, potentially leading to biased conclusions.

Recently, Rahman & Halim (2022) conducted a study to investigate the impact of gender on intrinsic, extrinsic, and STEM interests' careers utilizing a multivariate analysis of variance (MANOVA). Their findings revealed statistically significant differences in STEM career interests between male and female students, with males exhibiting higher STEM interests. However, no statistically significant impact of gender was observed on extrinsic and intrinsic determinants. Like previous studies, this research also utilized inappropriate statistical methods when analyzing ordinal-level Likert scale scores.

In summary, the previous studies conducted in Finland, Indonesia, and Malaysia have a typical limitation. The researchers frequently employed inappropriate statistical analysis methods when dealing with ordinal-level Likert scale data. This issue can lead to misinterpretation and erroneous conclusions, especially when comparing findings to those from more developed countries like Canada, and the USA. Addressing this limitation is vital for gaining a more accurate understanding of STEM education and career determinations.

Furthermore, it's important to recognize that the impact of these determinants on students' interest in STEM education and careers varies across regions and countries. This variability is well-supported by the review paper by Idris and Bacotang (2023). Determinants contributing to this variation include the availability of STEM resources in classrooms, ethnic diversity, national education policies, parental education and occupation, socioeconomic and cultural influences, students' self-efficacy, teaching and learning practices, and the competence of STEM teachers.

Additionally, there's a pressing need to bridge the knowledge gap in understanding the determinants that impact students' enrolment, performance, and retention in Additional Mathematics, a fundamental component of STEM education. Students with a strong foundation in mathematics are more likely to pursue STEM education and careers. Consequently, it's worth exploring the development of a robust statistical predictive learning-based algorithm based on these statistically significant determinants.

## **Research Methodology**

This section provided a brief overview of the research methodology in this study. Particularly, this study employed the CRISP-DM data science methodology, which was tailored to suit the study's requirements. In practice, there were six principal phases of the CRISP-DM data science methodology, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment, as detailed in the subsequent section. The principal reason for employing CRISP-DM data science methodology in this article was its robustness and reliable application in various industries (Authors, 2024; Tripathi et al., 2021).

# Business Understanding

In this study, the primary business objective is to investigate the determinants that significantly impact upper-secondary enrolment in Additional Mathematics, with a particular emphasis on the role of STEM human capital in fostering national economic growth. The aim is to increase the number of students pursuing STEM-related careers, which can have positive effects on job creation, unemployment reduction, poverty alleviation, and gender equality. Additionally, the data mining goal in this study is to develop a robust SESLA for predicting Additional Mathematics enrolment based on these determinants, following the Cross Industry Standard Process for data mining (CRISP-DM) data science methodology, as shown in Fig. 1.

However, this study acknowledges potential risks in this study, including the possibility of respondents providing inaccurate information due to the utilization of self-report questionnaires. Moreover, this study employed a one-stage cluster sampling technique to address the variations in educational offerings among secondary schools. Furthermore, this article also proposed a novel statistical method to monitor and mitigate underfitting and overfitting risks. Particularly, this study defined the absence of underfitting when the accuracy rates for both training and test sets are higher than 90%, which is based on the literature (Authors, 2013; Authors, 2017; Authors, 2023). Meanwhile, there is an absence of overfitting when no statistically significant findings in the Fisher's exact or Fisher-Freeman-Halton test. This study also highlighted that all programming in this study utilized free software, such as Microsoft Excel and R statistical software, running on a middle-end computing environment with an Intel(R) Core™ i5-10210U CPU@1.60GHz for cost-saving.

## Data Understanding

The data understanding phase served as the foundation for the business understanding phase, primarily focusing on identifying, collecting, and analyzing datasets to achieve the study's objectives. The targeted population in this study comprised urban upper-secondary students in the Kuantan District who were potential candidates for enrolling in Additional Mathematics. Notably, this study focused on the first batch of students who were enrolled in Additional Mathematics according to the KSSM syllabus.

Given the absence of a similar study and the related questionnaire in the existing literature, this study developed a new questionnaire by referencing relevant studies (Kae, 2010; Chong et al., 2014; Yao et al., 2016). The questionnaire's latent variables and determinants were based on the literature, encompassing two principal parts: Part A and Part B. Part A encompasses student enrolment in Additional Mathematics and socio-economic determinants, including the educational disciplines, ethnicity, family income, gender, parental education, and PT3 Mathematics achievements. Meanwhile, Part B encompasses intrinsic and extrinsic motivational determinants. The intrinsic motivational determinant, primarily self-efficacy was gauged across seven latent variables. In contrast, extrinsic motivational determinants such as parental influence, peer influence, and teacher influence were gauged utilizing five latent variables, respectively.



To ensure the internal reliability of the questionnaire, this research conducted a pilot study, distributing it to 50 randomly selected students and analyzing the collected data utilizing Cronbach's alpha. All Cronbach's alpha values for each determinant exceeded 0.7, confirming its reliability. It's important to note that this study did not perform a validity analysis through exploratory factor analysis (EFA) in conjunction with multiple correspondence analysis (MCA) for clustering (Authors, 2020), as the referred literature clusters latent variables based on the corresponding determinants.

The finalized questionnaire, based on the pilot study, was utilized in full-scale research conducted from mid-February to mid-March 2020 in Malaysia, before the first Movement Control Order (MCO) on March 18, 2020. The questionnaire was distributed to upper-secondary students from four randomly selected urban secondary schools in Kuantan District: *Air Putih* (3°49'42.6"N, 103°20'19.9"E), *Bukit Rangin* (3°48'7.8"N, 103°16'17.7"E), *Semambu* (3°52'12.0"N, 103°19'39.7"E), and *Tg Panglima Tg Muhammad* (3°49'18.1"N, 103°17'37.3"E). This study involved a total of 156, 115, 50, and 68 potential upper-secondary students enrolled in Additional Mathematics. It's important to note that this study employed a one-stage cluster sampling technique to minimize the variation in the educational package across these upper-secondary schools. Furthermore, the sample size utilized in this study met the minimum requirements, as determined by Yaman's equation, with a sampling error of 0.05 (Authors, 2021).

In the final stage of this phase, a screening test was conducted on the collected data from the full-scale research to uncover associations among qualitative determinants, including nominal-scale and ordinal-scale data. The Fisher's exact test and Fisher-Freeman-Halton test were employed for this purpose, selection over the chi-squared test utilized in the literature (Authors, 2021) due to their provision of exact test statistics. Furthermore, this study computed the likelihood of enrolment in Additional Mathematics utilizing the odds ratio and assessed its significance was determined utilizing the 95% confidence interval (C.I.) approach. Statistical significance was determined when the 95% C.I. of the odds ratio did not encompass the hypothesized value of the odds ratio equal to one.

## Data Preparation

Data preparation played a vital role in preparing the final dataset for modeling. This process involved addressing missing data, aggregating latent variables for determinants, recoding categorical determinants, integrating the dataset, and partitioning it into training and test sets utilizing appropriate ratios. Cross-tabulation, which was employed during the data understanding phase, aided in identifying incomplete respondents. In this study, the row deletion technique was selected to remove incomplete respondents, prioritizing data completeness and accuracy over MCA-based imputation techniques.

For determinants such as self-efficacy, parent influence, peer influence, and teacher influence, latent variables needed to be measured. This study aggregated these latent variables utilizing the median operator because arithmetic mean aggregation was deemed inappropriate for ordinal-scale variables, as highlighted in the literature review section. This aggregation process also facilitated the recording of nominal-scale and ordinal-scale variables with  $k$  categories into  $k - 1$  dummy variables. This article emphasizes that in qualitative research, failing to recode nominal-scale and ordinal-scale variables with  $k$

categories into  $k - 1$  dummy variables is inappropriate, as previously proposed and utilized in literature (Mohamad et al., 2016; Halid & Khalid, 2022).

The qualitative determinants were then integrated into a Common-Separated Values (CSV) dataset, which was randomly split into training and test sets. This article explored four distinct training-to-test ratios, namely 60:40, 70:30, 80:20, and 90:10, as the ideal ratio remained unknown. These ratios were evaluated utilizing innovative statistical techniques introduced in the modeling phase.

## Modeling

In accordance with the CRISP-DM data science methodology, the modeling phase aimed to predict the enrolment of urban upper-secondary students in Additional Mathematics within the Kuantan District. This study explored various statistical learning-based algorithms, encompassing both probabilistic and deterministic techniques, utilizing training sets. This study employed a probabilistic-based classification statistical learning-based algorithm, dichotomous logistic regression (benchmarks comparison), and deterministic-based classification algorithms, such as  $c$ -Support Vector Machine ( $c$ -SVM),  $\nu$ -SVM, C5.0, and Classification and Regression Tree (CART). Additionally, this study proposed four novel SESLAs: two dichotomous logistic regression- $c$ -SVM-based, and two dichotomous logistic regression- $\nu$ -SVM-based, with the principal objective of assessing their impact on classification accuracy rates, while the accuracy rates are evaluated based on the average of the diagonal elements of confusion matrices.

In this article, the primary objective of data mining was to develop a robust statistical predictive learning-based algorithm that not merely achieved a high accuracy rate ( $> 90\%$ ) (Authors, 2013; Authors, 2017; Authors, 2023), but also provided valuable insights for real-life applications. To maintain interpretability and manage computational cost, this article did not consider sophisticated artificial neural network (ANN) algorithms. Furthermore, ANOVA, MANOVA, MLR, panel regression, and SEM algorithms as proposed in the literature were also not relevant, given that this research was qualitative, and the data collected lacked time-dependent and spatial variation.

## Evaluation and Deployment

In the realm of technical algorithm evaluation, the evaluation phase assumes a pivotal role, guiding the deployment process based on specific requirements that can span from producing straightforward reports to implementing repeatable data mining procedures across enterprises. In this study, the test set was employed to assess the presence of overfitting, as well as to evaluate the robustness of probabilistic and deterministic classification statistical learning-based algorithms. Consequently, this study randomly generated 10 distinct scenarios by utilizing seeds ranging from 1 to 9999 and applied Wilcoxon's signed rank test to gauge the statistical significance of classification accuracy rates.

The analysis in this study aimed to determine the optimal training-to-test ratio by computing the absolute difference in accuracy rates between the training and test sets for each of these 10 scenarios. When the results did not demonstrate statistical significance, this indicated the robustness of the statistical predictive learning-based algorithm and the absence of overfitting. Notably, this study refrained from

relying on graphical techniques for assessing these issues, deeming them inadequate in providing statistical evidence. In summary, this innovative evaluation technique served not merely to establish optimal training-to-test ratios in cases of lacking prior knowledge but also to evaluate the robustness of statistical predictive learning-based algorithms in identifying superior classification techniques.

During the deployment phase, the superior statistical predictive learning-based algorithm was selected based on minimal accuracy rate differences between the training and test sets, along with an optimal training-to-test ratio. Specifically, the optimum training-to-test ratio is identified when there is an absence of underfitting and overfitting to ensure that the algorithm generalizes well. Nevertheless, the principal objective of this article was to solicit valuable feedback from experts in the field of science and mathematics education, with the aim of fostering continuous improvement in the future.

## **Analysis Results and Discussion**

This section outlines the statistical analysis utilized to achieve the study's principal objectives. Notably, this study employed Microsoft Excel and R statistical software for statistical analysis presented in this section. Additionally, the research methodology is primarily based on the CRISP-DM data science methodology, which is structured into five key phases: data understanding, data preparation, modeling, evaluation, and deployment.

## **Data Understanding and Data Preparation**

Table 1 presents the results of EDA for a comprehensive research dataset. Both Fisher's exact test and the Fisher-Freeman-Halton test revealed significant associations among several variables, including educational discipline, ethnicity, parental education, PT3 Mathematics achievements, mathematics self-efficacy, parental influence, peer influence, and teacher influence. These findings were specific to urban upper-secondary students in the Kuantan District enrolled in Additional Mathematics, without accounting for potential determinants and confounding effects.

Notably, students pursuing STEM packages were 44.32 times more likely to enroll in Additional Mathematics than those in Humanities & Arts packages, with a significant difference supported by a 95% C.I. of 21.87 to 89.78, excluding the hypothesized value of one. Similarly, students who achieved high PT3 Mathematics performance (Grade A and Grade B) were 142.73 times more likely to enroll in Additional Mathematics compared to those with low performance (Grade E and Grade F), supported by a 95% C.I. within 45.88 and 444.02. Conversely, students with average performance (Grade C and Grade D) were 14.44 times more likely to enroll in Additional Mathematics compared to those with low performance, with a 95% C.I. ranging from 5.02 to 41.54.

In line with previous studies, both intrinsic and extrinsic motivational determinants significantly impacted upper-secondary students' enrollment in Additional Mathematics. Students who strongly agreed or agreed (SAA) that mathematics self-efficacy had a strong impact were 117.14 times more likely to enroll in Additional Mathematics compared to those who strongly disagreed or disagreed (SDD), supported by a

95% C.I. within 38.77 and 353.93. Meanwhile, students who neither agreed nor disagreed (NAD) had a decreased likelihood, with mathematics self-efficacy having an impact of 20.34 times compared to those with SDD, with a 95% C.I. ranging from 11.24 to 36.80.

Additionally, parental influence, peer influence, and teacher influence are extrinsic motivational that are statistically significant on students enrolling in Additional Mathematics. Students who SAA that parental influence had an impact were 12.83 times more likely to enroll compared to those with SDD, with a 95% C.I. within 6.96 and 23.62. However, those who NAD had a 5.39 times higher likelihood compared to those with SDD, with a 95% C.I. ranging from 3.21 to 9.05. Likewise, students who SAA about the impact of peer influence were 24.57 times more likely to enroll in Additional Mathematics compared to those with SDD, with an interval estimate within 12.72 and 47.47. This was the most impactful extrinsic motivational determinant impacting student enrolment. Meanwhile, the likelihood of the students whose NAD believes the impact of peer influence has been reduced to 5.01 times more likely to enroll in Additional Mathematics compared to those who had SDD, resulting in a 95% C.I. lies within 2.86 and 8.76.

In contrast, teacher influence yielded a 15.79 higher likelihood when students with SAA, supported by a 95% C.I. ranging from 8.15 and 30.59. However, the likelihood dropped to 3.99 times with a 95% confidence interval within 2.34 and 6.80 when students merely NAD regarding teacher influence. Notably, determinants such as ethnicity and parental education did not have a statistically significant impact on students' enrollment in Additional Mathematics, as indicated by a 95% C.I. for odds ratios that included the hypothesized value of one. Likewise, no significant associations were found among the sub-categories of family income and gender, except for students from considerably high-income families (T20), who were 1.84 times more likely to enroll, supported by a 95% C.I. ranging from 1.07 and 3.16.

However, assessing the statistically significant associations and the likelihood of enrolment in Additional Mathematics without considering determinants and confounding effects provides an incomplete picture. Further analysis utilizing a statistical predictive learning-based algorithm revealed that the key determinants impacting student enrolment included educational disciplines, ethnicity, gender, mathematics self-efficacy, peer influence, and teacher influence, as detailed in the next section. As a result, this study did not delve into the practical implications of the statistically significant determinants presented in Table 1.

Table 1  
EDA on a Cleaned Full-scale Research Dataset

Determinant	Sub-categories	Symbolized	Frequency (Percentages)		Odds Ratio (95% C.I.)
			Enrolee	Non-enrolee	
Educational disciplines <sup>F*</sup>	STEM	$d_1$	168 (43.19%)	10 (2.57%)	44.32 [21.87, 89.78]*
	Humanities & Arts <sup>R</sup>	-	58 (14.91%)	153 (39.33%)	-
Ethnicity <sup>F*</sup>	Malay	$d_2$	94 (24.16%)	115 (29.56%)	0.73 [0.27, 1.96]
	Chinese	$d_3$	123 (31.62%)	40 (10.28%)	2.73 [0.99, 7.56]
	Indian & Others <sup>R</sup>	-	9 (2.31%)	8 (2.06%)	-
Family income	T20	$d_4$	60 (15.42%)	28 (7.20%)	1.84 [1.07, 3.16]*
	M40	$d_5$	74 (19.02%)	56 (14.40%)	1.13 [0.72, 1.80]
	B40 <sup>R</sup>	-	92 (19.02%)	79 (14.40%)	-
Gender	Male	$d_6$	94 (24.16%)	58 (14.91%)	1.29 [0.85, 1.95]
	Female <sup>R</sup>	-	132 (33.93%)	105 (26.99%)	-
Parental education <sup>F*</sup>	Tertiary	$d_7$	137 (35.22%)	73 (18.77%)	2.50 [0.84, 7.49]
	Secondary	$d_8$	83 (21.34%)	82 (21.08%)	1.35 [0.45, 4.06]
	Primary & Others <sup>R</sup>	-	6 (1.54%)	8 (2.06%)	-
PT3 Mathematics Achievement <sup>F*</sup>	High	$d_9$	145 (37.28%)	16 (4.11%)	142.73 [45.88, 444.02]*

*Note.* “F\*” indicates statistical significance to Fisher’s exact test or Fisher-Freeman-Halton test at a 0.05 significance level; “\*” indicates statistical significance at a 0.05 significance level.

	Average	$d_{10}$	77 (19.79%)	84 (21.59%)	14.44 [5.02, 41.54]*
	Low <sup>R</sup>	-	4 (1.03%)	63 (16.20%)	-
Mathematics Self- efficacy <sup>F*</sup>	Strongly Agree & Agree	$d_{11}$	80 (20.57%)	4 (1.03%)	117.14 [38.77, 353.93]*
	Neither Agree Nor Disagree	$d_{12}$	125 (32.13%)	36 (9.25%)	20.34 [11.24, 36.80]*
	Strongly Disagree & Disagree <sup>R</sup>	-	21 (5.40%)	123 (31.62%)	-
Parental Influence <sup>F*</sup>	Strongly Agree & Agree	$d_{13}$	97 (24.94%)	19 (4.88%)	12.83 [6.96, 23.62]*
	Neither Agree Nor Disagree	$d_{14}$	88 (22.62%)	41 (10.54%)	5.39 [3.21, 9.05]*
	Strongly Disagree & Disagree <sup>R</sup>	-	41 (10.54%)	103 (26.48%)	-
Peer Influence <sup>F*</sup>	Strongly Agree & Agree	$d_{15}$	125 (32.13%)	18 (4.63%)	24.57 [12.72, 47.47]*
	Neither Agree Nor Disagree	$d_{16}$	75 (19.28%)	53 (13.62%)	5.01 [2.86, 8.76]*
	Strongly Disagree & Disagree <sup>R</sup>	-	26 (6.68%)	92 (23.65%)	-
Teacher Influence <sup>F*</sup>	Strongly Agree & Agree	$d_{17}$	102 (26.22%)	18 (4.63%)	15.79 [8.15, 30.59]*
	Neither Agree Nor Disagree	$d_{18}$	96 (24.68%)	67 (17.22%)	3.99 [2.34, 6.80]*
	Strongly Disagree & Disagree <sup>R</sup>	-	28 (7.20%)	78 (20.05%)	-
<p><i>Note.</i> “F*” indicates statistical significance to Fisher’s exact test or Fisher-Freeman-Halton test at a 0.05 significance level; “*” indicates statistical significance at a 0.05 significance level.</p>					

## Modeling, Evaluation, and Deployment

This study undertook a robust data munging process to unveil hidden patterns within a meticulously cleaned full-scale research. This process harnessed the power of appropriate statistical tools, culminating

in the development of statistical predictive learning-based algorithms. The performance of these algorithms is detailed in Table 2. This study evaluated the predictive capabilities of one statistical probabilistic, and six statistical deterministic learning-based algorithms, in addition to four statistical stacked ensemble learning-based algorithms. This assessment considered various training-to-test ratios and included a benchmark comparison with the statistical probabilistic learning-based algorithm, dichotomous logistic regression, as found in previous studies. Additionally, this study also measured the average and standard deviation of the absolute distance between training and test accuracy rates, with the training accuracy rates determined through internal validation. These metrics provided valuable insights, offering a subjective gauge of the potential presence of overfitting within the employed competent subjectively to reflect the presence of overfitting issues on statistical predictive learning-based algorithms.

Furthermore, this article introduced an innovative approach for assessing both underfitting and overfitting. Specifically, this study identified the absence of underfitting when average accuracy rates for both the training and test sets exceeded 90%. This threshold holds significant recognition within the fields of pattern recognition and machine learning applications. Additionally, this study established the absence of overfitting when average accuracy rates for training and test sets yielded no significant differences, as confirmed by Wilcoxon's signed-rank test. These findings contribute to a more nuanced understanding of the robustness and generalization capabilities of statistical learning-based algorithms under investigation.

Table 2

Evaluating predictive performance and assessing the proposed statistical stacked ensemble learning-based algorithms

Training-to-test ratio	Algorithm	Accuracy rates		Absolute Distance mean (s.d)	Wilcoxon signed-rank test ( <i>P</i> -value)
		mean (s.d)			
		Training	Test		
60:40	Dichotomous logistic regression <sup>R</sup>	92.01 (1.56)	88.78 (2.51)	4.11 (2.46)	0.0322*
	CART	89.57 (1.11)	86.58 (2.73)	3.72 (2.84)	0.0414*
	C5.0	92.44 (1.76)	88.52 (3.19)	4.24 (2.94)	0.0144*
	c-SVM (RBF)	96.24 (1.02)	90.32 (2.49)	5.92 (3.35)	0.0059*
	$\nu$ – SVM (RBF)	93.42 (0.97)	90.19 (1.82)	3.23 (2.50)	0.0020*
	<b>c-SVM (Sigmoid)</b>	<b>90.04 (1.76)</b>	<b>90.52 (2.90)</b>	<b>3.24 (2.58)</b>	<b>0.9187</b>
	<b>\varvec{\nu} – SVM (Sigmoid)</b>	<b>90.47 (1.91)</b>	<b>90.52 (2.85)</b>	<b>3.26 (3.02)</b>	<b>0.7596</b>
	Stacked Logistic-c-SVM (RBF)	92.52 (1.64)	89.42 (2.47)	4.01 (2.58)	0.0488*
	Stacked Logistic- $\nu$ – SVM (RBF)	90.68 (1.75)	89.61 (2.48)	3.31 (2.04)	0.5406
	Stacked Logistic-c-SVM (Sigmoid)	89.10 (2.35)	88.00 (3.33)	3.38 (3.25)	0.3750
Stacked Logistic- $\nu$ – SVM (Sigmoid)	89.83 (2.77)	88.39 (3.88)	3.02 (2.90)	0.3223	
70:30	Dichotomous logistic regression <sup>R</sup>	91.83 (1.54)	89.05 (2.90)	3.75 (3.25)	0.1602
	CART	90.04 (1.16)	88.02 (3.84)	3.69 (2.54)	0.1260
	C5.0	92.68 (1.53)	88.88 (3.78)	4.52 (3.33)	0.0195*



Training-to-test ratio	Algorithm	Accuracy rates		Absolute Distance mean (s.d)	Wilcoxon signed-rank test ( <i>P</i> -value)
		mean (s.d)			
		Training	Test		
	c-SVM (RBF)	95.82 (0.88)	90.61 (3.19)	5.38 (3.65)	0.0098*
	$\nu$ – SVM (RBF)	92.64 (0.98)	90.26 (2.11)	2.89 (2.35)	0.0371*
	<b>c-SVM (Sigmoid)</b>	<b>90.10 (1.42)</b>	<b>90.43 (3.05)</b>	<b>3.42 (2.31)</b>	<b>0.9219</b>
	$\nu$ – SVM (Sigmoid)	90.70 (1.11)	89.91 (2.93)	3.05 (2.05)	0.5566
	Stacked Logistic-c-SVM (RBF)	92.56 (1.28)	89.22 (2.48)	4.00 (2.66)	0.0248*
	<b>Stacked Logistic- \varvec{\nu} – SVM (RBF)</b>	<b>90.88 (1.26)</b>	<b>90.09 (2.70)</b>	<b>3.06 (2.10)</b>	<b>0.6250</b>
	Stacked Logistic-c-SVM (Sigmoid)	89.19 (1.77)	87.59 (2.99)	3.64 (2.16)	0.3077
	Stacked Logistic- $\nu$ – SVM (Sigmoid)	90.62 (1.41)	89.40 (3.66)	3.88 (2.76)	0.6250
80:20	Dichotomous logistic regression <sup>R</sup>	91.80 (1.12)	89.35 (3.71)	3.83 (3.13)	0.1027
	CART	89.97 (0.74)	88.83 (2.95)	1.93 (2.16)	0.1525
	C5.0	92.31 (1.37)	90.00 (3.07)	3.09 (2.34)	0.0645
	c-SVM (RBF)	95.77 (0.72)	91.56 (2.55)	4.26 (2.97)	0.0080*
	<b>\varvec{\nu} – SVM (RBF)</b>	<b>92.18 (0.68)</b>	<b>90.78 (2.07)</b>	<b>2.31 (1.65)</b>	<b>0.1309</b>
	c-SVM (Sigmoid)	89.74 (1.04)	92.21 (2.94)	3.97 (1.99)	0.0664
	$\nu$ – SVM (Sigmoid)	<b>90.32 (0.83)</b>	<b>91.17 (3.11)</b>	<b>3.09 (1.97)</b>	<b>0.4145</b>

Training-to-test ratio	Algorithm	Accuracy rates		Absolute Distance mean (s.d)	Wilcoxon signed-rank test ( <i>P</i> -value)
		mean (s.d)			
		Training	Test		
	Stacked Logistic-c-SVM (RBF)	92.50 (1.25)	90.00 (3.18)	3.54 (3.25)	0.1260
	<b>Stacked Logistic- \varvec{\nu} – SVM (RBF)</b>	<b>91.09 (0.87)</b>	<b>90.78 (3.27)</b>	<b>3.70 (1.12)</b>	<b>0.9188</b>
	Stacked Logistic-c-SVM (Sigmoid)	88.69 (1.11)	88.70 (3.63)	2.51 (1.67)	0.8384
	Stacked Logistic- $\nu$ – SVM (Sigmoid)	89.90 (1.02)	90.13 (3.31)	2.99 (1.12)	0.8384
90:10	Dichotomous logistic regression <sup>R</sup>	91.71 (0.49)	89.74 (5.75)	5.26 (3.14)	0.4142
	CART	90.91 (0.64)	88.16 (4.85)	4.66 (3.26)	0.1025
	C5.0	92.39 (0.70)	89.21 (5.33)	4.85 (4.33)	0.1258
	c-SVM (RBF)	95.81 (0.33)	89.48 (6.33)	7.05 (5.58)	0.0246*
	$\nu$ – SVM (RBF)	91.97 (0.52)	88.95 (5.92)	5.65 (3.97)	0.2616
	<b>c-SVM (Sigmoid)</b>	<b>90.03 (0.76)</b>	<b>90.53 (4.51)</b>	<b>4.42 (2.37)</b>	<b>0.9187</b>
	<b>\varvec{\nu} – SVM (Sigmoid)</b>	<b>90.48 (0.65)</b>	<b>90.53 (4.68)</b>	<b>4.37 (2.55)</b>	<b>0.9187</b>
	Stacked Logistic-c-SVM (RBF)	92.65 (0.67)	88.95 (5.09)	5.32 (3.82)	0.0526
	<b>Stacked Logistic- \varvec{\nu} – SVM (RBF)</b>	<b>91.68 (0.48)</b>	<b>90.79 (4.68)</b>	<b>3.67 (3.15)</b>	<b>0.9188</b>
	Stacked Logistic-c-SVM (Sigmoid)	88.92 (1.02)	89.74 (5.18)	3.91 (3.54)	0.6101

Training-to-test ratio	Algorithm	Accuracy rates		Absolute Distance mean (s.d)	Wilcoxon signed-rank test ( <i>P</i> -value)
		mean (s.d)			
		Training	Test		
	Stacked Logistic- $\nu$ – SVM (Sigmoid)	91.08 (0.67)	90.00 (4.93)	4.40 (3.16)	0.6250

In Table 2, the analysis results revealed that statistical decision-tree-based learning-based algorithms, such as CART and C5.0, performed poorly across various training-to-test ratios. These algorithms failed to meet the underfitting and overfitting criteria, especially at the 60:40 training-to-test ratio. As a result, this study concluded that statistical decision-tree learning-based algorithms were not suitable for predicting the enrolment of urban upper-secondary students in Additional Mathematics, despite their automatic determinants selection ability.

Furthermore, when employing a 60:40 training-to-test ratio, merely c-SVM (accuracy rates-training set: 90.04%, test set: 90.52%; *P*-value for Wilcoxon signed-rank test: 0.9187) and  $\nu$ -SVM (accuracy rates-training set: 90.47%, test set: 90.52%; *P*-value for Wilcoxon signed-rank test: 0.7596), both utilizing the sigmoid kernel function, were identified as superior statistical predictive learning-based algorithms without encountering underfitting or overfitting issues. However, these algorithms for all levels of training-to-test ratio were not designated as superior statistical predictive learning-based algorithms due to their lack of interpretability and inability to select statistically significant determinants. Achieving interpretability for the determinants selected by these algorithms had proved challenging, and they had been unable to provide statistical evidence for the determinants when compared to the benchmark algorithm.

To address the limitations of statistical SVM-based learning-based algorithms, this study was motivated to develop a stacked ensemble statistical learning-based algorithm by combining dichotomous logistic regression and statistical SVM learning-based algorithms. Specifically, dichotomous logistic regression played a crucial role in selecting a set of statistically significant determinants utilizing stepwise selection techniques. Consequently, the selected statistically significant determinants were utilized for prediction based on statistical SVM learning-based algorithms, resulting in a complementary stacked ensemble statistical learning-based algorithm.

Contrarily to the SVM-based utilizing sigmoid kernel function, the analysis results revealed that the proposed stacked ensemble statistical learning-based algorithms utilizing sigmoid kernel function, including stacked dichotomous logistics regression-c-SVM and dichotomous logistics regression- $\nu$ -SVM did not meet neither underfitting nor overfitting. Despite that, Table 2 revealed that the stacked dichotomous logistics regression- $\nu$ -SVM utilizing the well-recognized gaussian radial basis function (RBF) kernel function is outperformed compared to other algorithms across 70:30, 80:20, and 90:10 training-to-test ratios. This analysis results also lead to the superior statistical predictive learning-based

algorithm in predicting the urban upper-secondary students to enroll in Additional Mathematics is stacked dichotomous logistics regression- $\nu$ -SVM algorithm with utilizing RBF kernel function, while the optimal training-to-test ratio is 70:30, 80:20, and 90:10.

Furthermore, the analysis results revealed the absolute distance accuracy rates between training and test sets for the 70:30, 80:20, and 90:10 ratios, based on 10 random seeds. The absolute distance accuracy rates ranged from 0.69–6.60%, 1.50–5.30%, and 0.09–10.44%, respectively. However, the Wilcoxon signed-ranked test revealed no statistical differences between the accuracy rates for training and test sets. Thus, this study selected the significant determinants of the stacked ensemble algorithm, choosing those with the least absolute distance accuracy rate between training and test sets, and which met the underfitting and overfitting criteria. In other words, the superior significant determinants were acquired from the stacked ensemble statistical learning algorithm trained utilizing an 80:20 training-to-test ratio, with an absolute distance of 1.50%. These statistically significant determinants include educational discipline, ethnicity, gender, mathematics self-efficacy, peer influence, and teacher influence. In the context of Malaysia's education system, Additional Mathematics is integral to the STEM educational discipline, especially for students specializing in pure science subjects (Radhi & Arumugam, 2020). Consequently, students pursuing STEM educational disciplines are more likely to enroll in Additional Mathematics compared to their counterparts in Humanities & Arts educational disciplines.

Moreover, the analysis results revealed that ethnicity is a significant determinant impacting upper-secondary students' enrolment in Additional Mathematics. Specifically, the findings indicated that Malay ethnicity is associated with a lower likelihood of enrolling in Additional Mathematics compared to the Indian & Others ethnic groups, while no statistically significant difference was observed for the Chinese ethnic group. In Kuantan District, where approximately 80% of the population is of Malay ethnicity, a substantial proportion of respondents (53.72%) in this study also identified as Malay. Notably, the distinct enrolment patterns, with 29.56% of Malay ethnic students choosing not to enroll, compared to only 2.06% in the Indian & Others ethnic groups, as presented in Table 1, underscore the significance of ethnicity as a determinant within the logistic regression algorithm. These findings are consistent with international research, such as a study by Niu (2017), and Idris and Bacotang (2023), which similarly recognized ethnicity as a significant determinant in the context of student enrolment.

In practice, male students were found to be more likely to enroll in Additional Mathematics compared to their female counterparts, aligning with existing literature (Niu, 2017; Bowden et al., 2018; Panizzon et al., 2018; Kaleva et al., 2019). This observation is consistent with studies indicating that male students tend to pursue higher-paying STEM fields such as engineering, physical sciences, mathematics, and computer sciences, while female students exhibit lower mathematical aptitude, self-concept, and interest in STEM-related subjects.

Similarly, the particular intrinsic and extrinsic motivational determinants as depicted in Table 1, including mathematics self-efficacy, peer influence, and teacher influence, significantly impacted students' decisions to enroll in Additional Mathematics. Students who SAA and NAD believed in the impact of

mathematics self-efficacy were more likely to enroll compared to those who SDD. This aligns with the concept of self-efficacy, reflecting students' beliefs in their capability to perform necessary behaviors for mathematics achievement (Bandura, 1977).

Teachers and counselors play a pivotal role in guiding students toward STEM careers and promoting gender equality in STEM fields. This analysis results showed that students who SAA believed in the influence of teachers and peers were more likely to enroll in Additional Mathematics than those students who SDD. These findings emphasized the importance of empowering educators and counselors to guide students in STEM career paths though in the digital economy and promote gender equality in STEM fields through curriculum enhancements, adjusted teacher expectations, improved educational tracking, and supportive peer interactions.

The study's insights are valuable for educational policymakers, highlighting the significance of STEM education initiatives. Policymakers can focus on empowering educators and counselors to guide students toward STEM careers and promoting gender equality in STEM fields through various measures. This study's findings were disseminated in a research article, serving as a report to gather feedback from experts in educational studies.

## **Conclusions and Future Work**

In conclusion, this study has successfully identified significant determinants impacting urban upper-secondary students' decisions to enroll in Additional Mathematics within the Kuantan District. The determinants found to be influential include education disciplines, ethnicity, gender, mathematics self-efficacy, peer influence, and teacher influence. Additionally, the introduction of a novel stacked ensemble statistical learning-based algorithm has improved predictive accuracy compared to traditional dichotomous logistic regression algorithms. These findings offered vital insights for educational policy and practice, emphasizing the importance of promoting STEM education initiatives. The finding of this study encourages the empowerment of educators and counselors to guide students toward STEM careers and to work on promoting gender equality within STEM fields. Furthermore, this study suggests that future research should include a comparative analysis between the resulting superior predictive algorithm and stacked Bayesian-based statistical algorithms and apply post-technique of Platt scaling to the resulting superior statistical predictive learning-based algorithm in this article, aiming to provide a more comprehensive understanding of the determinants impacting upper-secondary students' decisions to enroll in Additional Mathematics. This study contributes significantly to the field, offering valuable guidance for policymakers and educators.

## **Declarations**

### **Declaration of Conflict of Interest**

The author declared that there is no conflict of interest.

## Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT to improve the readability and language of this work. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

## Awards

The statistical learning-based algorithm, specifically the dichotomous logistic regression algorithm described in this article, was awarded bronze medals in the 2020 International Educational Invention, Innovation & Design Competition (iVEDIIC2020) within the innovation categories, and in the 2021 International Multidisciplinary Innovation Competition (IMIC2021) within the academic and professional social science categories under the innovation without production sector. Additionally, our innovation received a special award for the best extended abstract in IMIC2021.

## References

1. Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191-215. <https://doi.org/10.1037/0033-295X.84.2.191>
2. Bernama. (2021, October 9). Malaysia needs to increase percentage of students in STEM-Dr. Adham. <https://www.nst.com.my/news/nation/2021/10/734985/malaysia-needs-increase-percentage-students-stem-dr-adham>
3. Blotnicky, K. A., Franz-Odendaal, T., French, F., & Joy, P. (2018). A study of the correlation between STEM career knowledge, mathematics self-efficacy, career interests, and career activities on the likelihood of pursuing a STEM career among middle school students. *International Journal of STEM Education*, 5, 22. <https://doi.org/10.1186/s40594-018-0118-3>
4. Bowden, M., Bartkowski, J. P., Xu, X., & Jr., R. L. (2018). Parental occupation and the gender math gap: examining the social reproduction of academic advantage among elementary and middle school students. *Social Sciences*, 7(1), 6. <https://doi.org/10.3390/socsci7010006>
5. Chong, C. C., Lin, L. W., Chuen, L. C., Chai, T. T., & Yi, Y. W. (2014). *A study on factors influencing students' intention to pursue higher education* [Unpublished bachelor research project]. Universiti Tunku Abdul Rahman.
6. Authors (2013). *Sains Malaysiana*.
7. Authors (2017). *Journal of Physics: Conference Series*.
8. Authors (2021). *Journal of Physics: Conference Series*.
9. Authors (2024). *AIP Conference Proceedings* [Accepted for Publication].
10. Authors (2023). *Malaysian Journal of Fundamental and Applied Sciences*.
11. Department of Statistics Malaysia. (2023, September 7). *Labour Force Survey (LFS) time series statistics by state, 1982-2022: employed persons by industry, Malaysia/states, 1982-2022*.

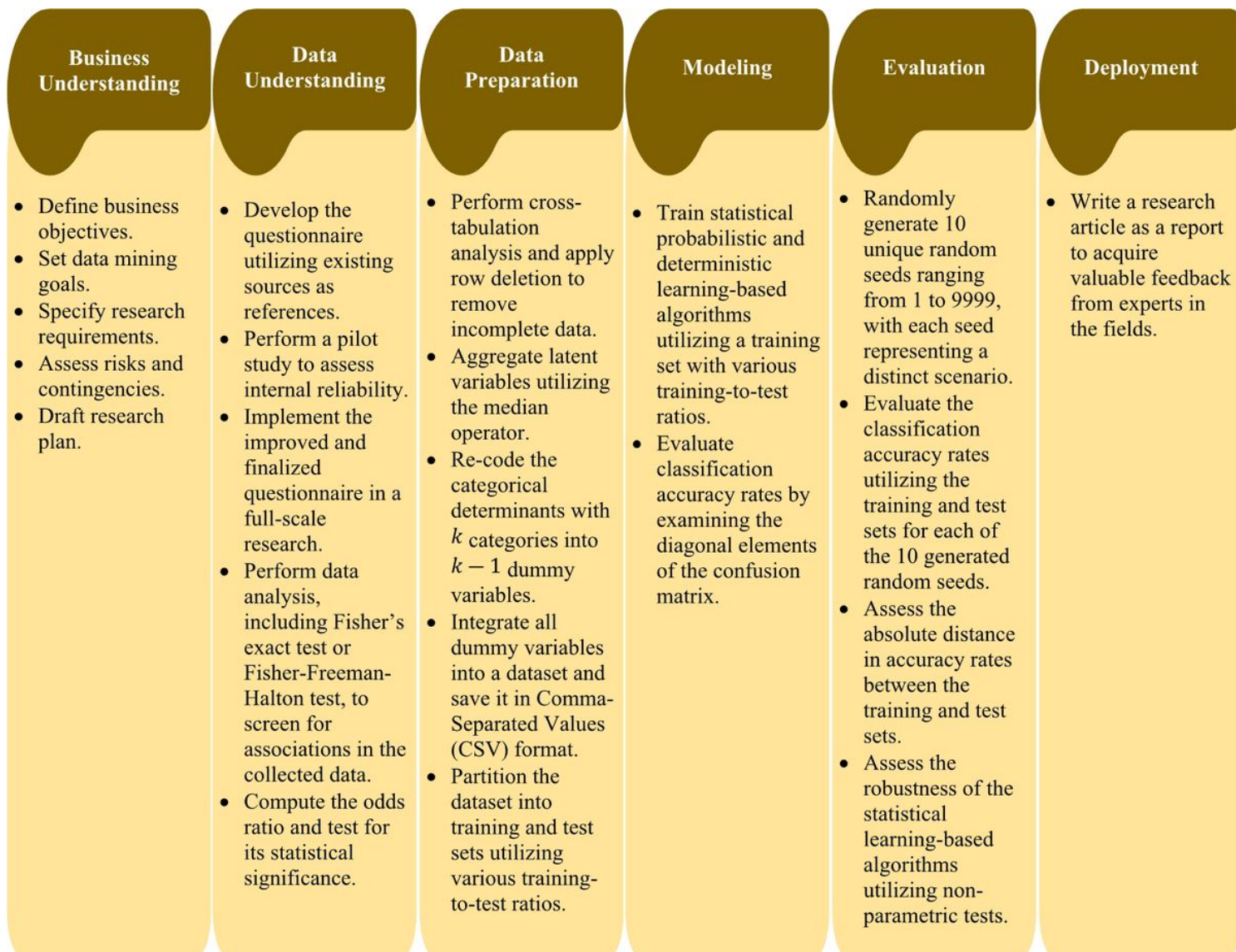
<https://www.dosm.gov.my/portal-main/time-series>

12. Dom, T. S. A. M. (2019, December 19). *KSSM perkasakan pelajar secara holistik*. <https://www.bharian.com.my/kolumnis/2019/12/636767/kssm-perkasakan-pelajar-secara-holistik>
13. Halid, N. H. M., & Khalid, Z. M. (2022). Modeling graduate employability in Malaysia using logistic regression. *Proceedings of Science and Mathematics, 9*, 1-11.
14. Halim, L., Rahman, N. A., Ramli, N. A. M., & Mohtar, L. E. (2018). Influence of students' STEM self-efficacy on STEM and physics career choice. *AIP Conference Proceedings, 1923*(1), 020001. <https://doi.org/10.1063/1.5019490>
15. Hiae, T. E. (2020). *Logistic regression analysis in determining the enrolment likelihood of Additional Mathematics for Kuantan District* (Unpublished Master's Dissertation). Universiti Malaysia Pahang.
16. Idris, R., & Bacotang, J. (2023). Exploring STEM education trends in Malaysia: building a talent pool for Industrial Revolution 4.0 and Society 5.0. *International Journal of Academic Research in Progressive Education & Development, 12*(2), 381-393. <http://dx.doi.org/10.6007/IJARPED/v12-i2/16825>
17. Kae, W. J. (2010). *Faktor-faktor yang mempengaruhi pembelajaran Matematik Tambahan dalam kalangan pelajar tingkatan empat* [Unpublished bachelor final year project report]. Universiti Teknologi Malaysia.
18. Kaleva, S., Pursiainen, J., Hakola, M., Rusanen, J., & Muukkonen, H. (2019). Students' reasons for STEM choices and the relationship of mathematics choice to university admission. *International Journal of STEM Education, 6*, 43. <https://doi.org/10.1186/s40594-019-0196-x>
19. Kamsi, N. S., Firdaus, R. B. R., Razak, F. D. A., & Siregar, M. R. (2019). Realizing Industry 4.0 through STEM education: but why STEM is not preferred?. *IOP Conference Series: Materials Science and Engineering, 506*, 012005. <https://doi.org/10.1088/1757-899X/506/1/012005>
20. Malay Mail. (2016, December 31). *New KSSM, KSSR curriculum to be implemented in 2017, says education minister*. <https://www.malaymail.com/news/malaysia/2016/12/31/new-kssm-kssr-curriculum-to-be-implemented-in-2017-says-education-minister/1283505>
21. Ministry of Education Malaysia. (2019). *Kurikulum Standard Sekolah Menengah Matematik Tambahan: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 4 dan 5 (Edisi Bahasa Inggeris)*. <http://bpk.moe.gov.my/index.php/terbitan-bpk/kurikulum-sekolah-menengah/category/397-form-4-and-5>
22. Mohamad, N. A., Ali, Z., Noor, N. M., & Baharum, A. (2016). Multinomial logistic regression modelling of stress level among secondary school teachers in Kubang Pasu District, Kedah. *AIP Conference Proceedings, 1750*(1), 060018. <https://doi.org/10.1063/1.4954623>
23. Mohtar, L. E., Halim, L., Rahman, N. A., Maat, S. M., Iksan, Z. H., & Osman, K. (2019). A model of interest in STEM careers among secondary school students. *Journal of Baltic Science Education, 18*(3), 404-416. <https://doi.org/10.33225/jbse/19.18.404>
24. Niu, L. (2017). Family socioeconomic status and choice of STEM major in college: an analysis of a national sample. *College Student Journal, 51*(2), 298-312.

25. Panizzon, D. L., Geer, R., Paige, K., O'Keeffe, L., Schultz, L., Zeegers, Y., & Brown, L. (2018). Exploring the 'hard facts' around STEM in Australia: females, low socioeconomic status and absenteeism. *International Journal of Innovation in Science and Mathematics Education*, 26(8), 30-44.
26. Radhi, N.A.M., & Arumugam, T. (2020, January 5). *Rough start for new upper secondary curriculum*. <https://www.nst.com.my/news/nation/2020/01/553638/rough-start-new-uppersecondarycurriculum#:~:text=For%20the%20STEM%20package%2C%20students,%3B%20or%20at%20least%20two>
27. Rahman, N. A., & Halim, L. (2022). STEM career interest: the effect of gender. *Creative Education*, 13(8), 2530-2543. <https://doi.org/10.4236/ce.2022.138160>
28. Ramli, N. A. M., & Awang, M. (2020). Critical factors that contribute to the implementation of the STEM education policy. *International Journal of Academic Research in Business & Social Sciences*, 10(1), 111-125. <http://doi.org/10.6007/IJARBSS/v10-i1/6811>
29. Razali, F., Talib, O., Manaf, U. K. A., & Hassan, S. A. (2017). A measure of students motivation, attitude and parental influence towards interest in STEM career among Malaysian form four science stream student. *International Journal of Academic Research in Business and Social Sciences*, 7, 245-264. <https://doi.org/10.6007/IJARBSS/v7-i14/3665>
30. Authors (2020). *Malaysian Journal of Mathematical Sciences*.
31. Shahali, E. H. M., Halim, L., Rasul, M. S., Osman, K., & Zulkifeli, M. A. (2017). STEM learning through engineering design: impact on middle secondary students' interest towards STEM. *EURASIA Journal of Mathematics Science and Technology Education*, 13(5), 1189-1211. <https://doi.org/10.12973/eurasia.2017.00667a>
32. The Vibes. (2023, July 20). *Mosti to raise awareness on science, technology, innovation as enrollment drops*. <https://www.thevibes.com/articles/news/96585/mosti-to-raise-awareness-on-science-technology-innovation-as-enrollment-drops#:~:text=He%20said%20that%20based%20on,from%2045.20%25%20to%2040.94%2>
33. Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in Artificial Intelligence*, 4, 576892. <https://doi.org/10.3389/frai.2021.576892>
34. United States Department of Labor. (2023, September 6). *Employment projections*. <https://www.bls.gov/emp/tables/stem-employment.htm>
35. Wen, H. J. (2023, June 18). *STEM-ming the decline*. <https://www.thestar.com.my/news/education/2023/06/18/stem-ming-the-decline>
36. Yao, A. M., Hui, L. S., Xin, N. Y., Chen, T. K., & Wei, W. K. (2016). *Factors affecting students to enrol in finance related major* [Bachelor Research Project]. Universiti Tunku Abdul Rahman.

## Figures





**Figure 1**

Schematic of CRISP-DM data science methodology: application in STEM educational study