Pakistan Journal of Life and Social Sciences
www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.00562

**RESEARCH ARTICLE**

# Educational Big Data Analytics: Machine Learning Based Academic Performance Predictive Modelling

Ng Yi Ling[1*], Ting Tin Tin[2], Teoh Chong Keat[3], Umar Farooq Khattak[4], Mohammed Amin Almaiah[5]

[1,2] Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan

[3] Artificial Intelligence Research and Computational Optimization (AIRCO) Laboratory, DigiPen Institute of Technology Singapore

[4] School of IT, Unitar International University, Malaysia

[5] Department of Computer Networks and Communications, College of Computer Science and Information Technology, King Faisal University, Al-Ahsa 31982, Saudi Arabia

***ARTICLE INFO***

Received: Aug 22, 2024

Accepted: Oct 27, 2024

***Keywords***

Education Data Mining

Academic Performance

Star Schema

Imbalance Dataset

Resampling

Hyperparameter Tunning

SMOTE

***Corresponding Author**

ylng021988@hotmail.com

**ABSTRACT**

Education is the key to success for a country moving forward to a country of sustainable development. This has been highlighted as one of the Sustainable Development Goals, which is Quality Education. Quality education is highly dependent on the services provided by educational organisations, which are reflected in the student's academic performance. In this research, the PRISMA method is used to perform a review of the literature on the factors that impact student academic performance and its associated predictive models. Most previous research on extracting educational data is based on a small population and limited factors. A larger population needs to be analysed to ensure that educational organisations can implement a comprehensive and generalized model to predict students' academic performance without bias. The Star schema combines three datasets, generating two million records distributed into four academic performance categories: Excellent, Very Good Good, and Fair. The issue of unequally distributed records among the four categories is solved using the Synthetic Minority Oversampling Technique (SMOTE) to produce 420,972 records in each academic performance category. Predictive modelling using multiple algorithms (Generalised Linear Model, Random Forest, Decision Tree, Fast Large Margin, Logistic Regression, Naïve Bayes, Support Vector Machines, Deep Learning) results in deep learning that outperforms other algorithms, with 90.1 precision and 9.9 classification errors. The deep learning model has been chosen to improve its performance further using a random search method and this has successfully increased the accuracy to reach 100%. The prepared model can predict the student's academic performance; after knowing that the student is believed to be underperforming, action should be taken to help the student perform better. Therefore, a report on the root causes that affect student academic performance must be generated to propose corrective action to help students, which is part of the role of educational organisations. This can be part of the service an educational organisation is doing to improve its service of providing quality education.

## 1. INTRODUCTION

United Nations has listed seventeen Sustainable Development Goals, where quality education is among the most important goals (United Nations, 2023). The educational organisation provides services that allow students to learn and understand new skills. This is also a place that brings all the

learners together to help each other in learning. Therefore, improving the quality of the educational service is very important to ensure quality education is given to the students. Previous studies have found many measurements used to measure an educational organisation's quality. Some educational organisations use graduation rates or dropout rates (Rajendran et al., 2022), and some use the Grade Point Average (GPA) as the benchmark (Ojajuni et al., 2021). No matter which way of measurement, all educational organisations are heading toward the same goal: to make sure students can understand the world within themselves and around them (Robinson & Robinson, 2022).

All educational organisations are heading to the following four main educational objectives: personal, cultural, economic, and social. From a unique perspective, this includes building a person's confidence level by making the person's dream come true. When a person is skilful and knowledgeable, they can plan for their future. Whenever all the planning has come to a realisation, the person's confidence level will increase accordingly. Besides, a person with knowledge can protect himself. A knowledgeable person should know how to read and write and have basic analytic skills, which will reduce the chance of being scammed. On the other hand, education will have self-awareness in knowing their strengths and weaknesses (Ashokram, 2021). Education helps to build a person's personality, including interpersonal skills and cultivation of capabilities and aptitude. With the skills and knowledge that a person learns from the educational organisation, a person can get a better career. The person can secure a higher pay job with the certification that the educational organisation gives after their learning. Education can also help them achieve sustainable global goals: no poverty and zero hunger (United Nations, 2023).

Secondly, education plays an essential role in culture. Different race groups have their history, making them have different cultures and beliefs. Each has wisdom and value of existence (Pusat Bahasa Abad, 2018). A proverb in Bahasa Malaysia says, “Tak kenal maka tak cinta,” which means not knowing hence not loving (Yusof, 2013). To fall in love with their or others' culture, they must know the ins and outs of all these cultures. That is when a person can accept and contribute to the culture (Robinson & Robinson, 2022). Education is bridging everyone together by reducing the gaps in society. At the same time, education helps to reduce inequality among society and country and maintain gender equality. Both are the goals of the United Nations moving towards sustainable development (United Nations, 2023).

Following culture, Economics is the third main objective in education. Education is an avenue for developing a country (Ebenuwa-Okoh, 2010). An educated society is needed to boost a country's economic growth. When a country has a majority of the population with higher education, more employment opportunities will be opened. This will help to increase the productivity of the country. When most of the population has a stable income, this will lead to steady economic growth (Natural Beach Living, 2023). Research carried out in the United States shows that when the number of graduates increases, the country's Gross Domestic Product (GDP), annual earnings, and expenses will also increase; this indirectly causes the Federal Tax Revenue to grow (Anna Sudderth, 2022). With a highly educated population, a country can help improve the efficiency of its workforce (Grant, 2017). We can conclude that a well-educated generation will be able to lead the country marching towards development in all aspects (S. Li & Liu, 2021).

Lastly, education also helps in society. Reports on “A Human Approach to World Peace” mentioned that misunderstanding is the primary human problem. This can be resolved by education (Dalailama, 2023). Training a person to have better interpersonal skills will reduce the loss of miscommunication that happens during work and daily life (Natural Beach Living, 2023). This will create a more peaceful and harmonious society. Moral values have to be fostered in youth to have good behaviour and attitude, which will cultivate a generation aware of their responsibility to create a peaceful society. This will lead the country towards peace, justice, and a vital institution. This aligns with the sixteenth sustainable development goal (United Nations, 2023). With a more knowledgeable population, not only is the well-being of the world of the country taken care of, but the inner state of the country's population will also increase. Education is the solution, allowing the government to reach good health and well-being and fulfil one of the goals of a sustainable country (United Nations, 2023).

Education is a key to enabling a country to move toward a sustainable nation and world. An educational organisation must improve their service to increase the graduation rate. Before any action can be taken to improve the educational service to students, the educational organisation needs to understand the factors affecting students' academic performance. There are two ways researchers find out the factor affecting student's performance. 1) The researcher finds out the factor using statistical analysis. 2) Researchers find out the factor using data mining. In this research, educational data mining (EDM) will focus on determining the factors affecting students' performance and preparing a Machine Learning (ML) model to predict students' academic performance. With the prediction of academic performance, educational organisations can take precautionary steps to prevent students from dropping out or failing to graduate (Cagliero et al., 2021).

The research done by Agnafors et al. (2021) concludes that mental health problems that occur in early childhood will increase the high-risk student do not perform well during their academic year. Different genders at different age levels will be one of the factors affecting the student's academic performance. Briones et al. (2021) also highlighted that students' attitudes going to school, such as lack of motivation, can affect academic performance (Briones et al., 2021). Student attitudes toward study, such as being lazy to study, studying when needed, having no time to learn at home, and being disturbed when studying, can be one factor that affects the student's academic performance (Olufemioladebinu et al., 2018). Above are the factors related to students found in the previous statistical analysis research.

Some factors can be categorised into family or parents-related factors that affect students' academic performance. David Garcia-Gonzalez et al. (2019) found parents' educational level, socioeconomic status, residence area, number of rooms, parent annual income, and predominant material at home. Besides, the number of siblings and the number of people residing in the same house unit are the factors which are family related that can affect the student's academic performance (García & Skrita, 2019). Research by Li & Qiu (2018) aligns with research from Gracia & Skrita (2019), as both mentioned that the family's socioeconomic status would impact students' academic performance. Parent background is being highlighted in Olufemioladebinu et al. (2018) research. Parents' attitudes toward students' studies, like supportive, democratic, and strict, could affect students' academic performance (Briones et al., 2021). This has clearly shown that family members and the environment, at some point, will make some changes to students' academic performance.

The educator or teacher is the person who is responsible for transmitting knowledge to students in an educational organisation. Xi Jinping, the general secretary of the Chinese Communist Party, used to call a teacher "The engineers of the human soul". What teachers spread in the educational organisation will shape the soul and life of students. They also spread ideas and truths other than knowledge in the educational organisation (Wikipedia, 2022). Arora & Singh (2017) mentioned that how teachers plan the class to make it a practical class carries the highest variance among all the factors. All the factors mentioned above are categorised into factors related to teachers.

Lastly, the educational organisation provides an environment to transmit knowledge from educators to students. Quality of the resources in school, inclusive of the experiment lab, quality of books available in the library, condition of the classroom, and equipment available in the school, can be the factors that affect the student's academic performance (Z. Li & Qiu, 2018). The bad experiences that happened in educational organisations, like the deadliest incident that occurred in an elementary school in Newtown, US (M. Ray, 2023), will cause students to have early behavioural problems (Agnafors et al., 2021). Statistics show that students who go through gun violence will be depressed, which will then lead to absenteeism and a decline in academic performance (Rossin-Slater, 2022). In this research, all the factors will be categorised into four main categories: Student-related, family/parent-related, teacher-related, and school-related.

Previous research is quite comprehensive in meeting the objective of determining the factors affecting students' academic performance. The research can be improved by comparing the factors

from all aspects instead of focusing on one or two categories. Some research uses a big enough dataset and concentrates only on mathematical achievement (Bayirli et al., 2023). Huynh-Cam et al. (2022) suggest that more factors should be put into the prediction model to allow better prediction. This is in line with Sayed et al. (2023) and Lopez-Zambrano et al. (2021), which highly recommended that multiple datasets be put into consideration when preparing the predictive model. This enables the model to develop a general framework suitable for implementation at all educational organisation levels (López-Zambrano et al., 2021). PISA 2018 is a big dataset (PISA, 2020). The PISA 2018 dataset contains responses from 79 countries with 120,000 students and their parents, 107,000 teachers, and 21,900 school principals (PISA, 2020). This tall and wide dataset is eligible to help prepare a generalised model that can be implemented in different countries.

With the help of the PISA dataset, the underlying factor that affects the student's academic performance can be uncovered. Educational organisations will be able to understand students' possible patterns and behaviour in this era. This will help the educational organisation prepare the course and mode of teaching suitable for the students. This will also help the stakeholders of the educational organisation make decisions and plan according to the student's needs. This can help educational organisations improve the service that is provided. This is an academic service and other value-added services, such as libraries, laboratories, and the hostel's environment. Recognising the students' weak points early can help reduce the failing rate. Precaution steps can be taken to avoid student withdrawal from courses (Rawat, 2022). Big data can give educational organisations a deep insight into students' academic performance and help them make decisions based on the trends found (Polymer, 2023). When trained with an extensive dataset, the model will learn different combinations of factors that affect students' academic performance. Therefore, during prediction, the model will produce more accurate results.

## 2. PRISMA Model in systematic literature review

The overview process of this systematic literature review includes five steps: 1. Define a topic, 2. Develop research objective, 3. Keywords identified, 4. Article identified and searched, and 5. Read and access the publication. The defined research title is used to develop five scopes to be focused on in this research as shown in Table 1. Keywords are identified based on the topics.

**Table 1: Research title with associating research topic and identified keywords**

| **Research title** | **Research topics (Corresponding Section)** | **Keywords** |
|---|---|---|
| Educational big data analytics | Predictors for academic performance (Section 3) | Factor impacting AP<br>Teacher impacting AP<br>Parents impacting AP<br>School impacting AP<br>Student attitude affecting AP |
| | Data modeling (Section 4) | What is data modelling<br>Why data modelling important<br>Star schema<br>Snowflake schema<br>Usage of schema |
| | Imbalance classification (Section 5) | Drawback of imbalance classification<br>How to overcome imbalance classification<br>SMOTE |
| Machine learning-based academic performance predictive modeling | Machine learning-based academic performance prediction research (Section 6) | Machine learning AP<br>Forecasting student performance<br>Data mining student performance<br>Model to predict student performance<br>Educational data mining |

| | | |
|---|---|---|
| | Hyperparameter tuning and model testing<br>(Section 7) | Artificial intelligence model for student performance prediction<br><br>Hyperparameter tuning<br>Model testing<br>Advantage of the hyperparameter tunning |

Note: AP - academic performance

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) is the chosen literature review model for this research. PRISMA Model is a model which is flexible to apply in any type of research to assist the researchers in organising collected information. PRISMA model will help researchers understand the main objective of doing the research, scopes that have been covered by previous researchers from the same field and knowledge that was learnt from previous research. The updated version of PRISMA 2020 proposed researchers take the three main steps while doing a literature review which are identify, screen and include. During each step, the researchers must go through an elimination process before proceeding to the next step. The eliminating process is to remove the duplicate record or the record which is not relevant to the research objective. The researcher will first identify the research paper from databases, registries and other resources like websites. Then, screening through the article by cross-checking it with the research objective. Lastly, the researcher will decide if the article is to be included as part of the literature review (Page et al., 2021). Figure 1 is the PRISMA Model implemented in this research study.

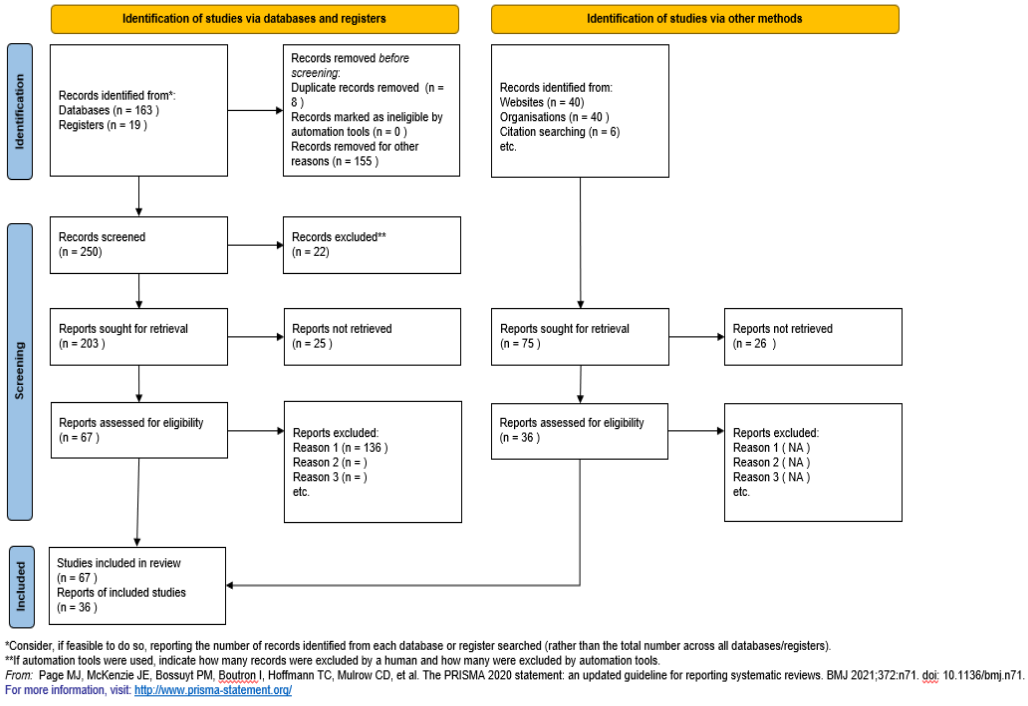


**Figure 1: PRISMA model**

The article found was from one hundred and eighty-two different databases and registries. Around eight duplicated records are removed from the list. Around one hundred fifty-five other records were removed from the list due to the article already aged more than five years. Two hundred fifty records are being screened. On the other hand, there are around forty-six websites and citation searching visited. Only seventy-five reports were retrieved and a third-six of them are eligible to be included in this study. The eligible database, registries and website are shown in Figure 2. Figure 3 is the list of databases and registries which are used in this study.

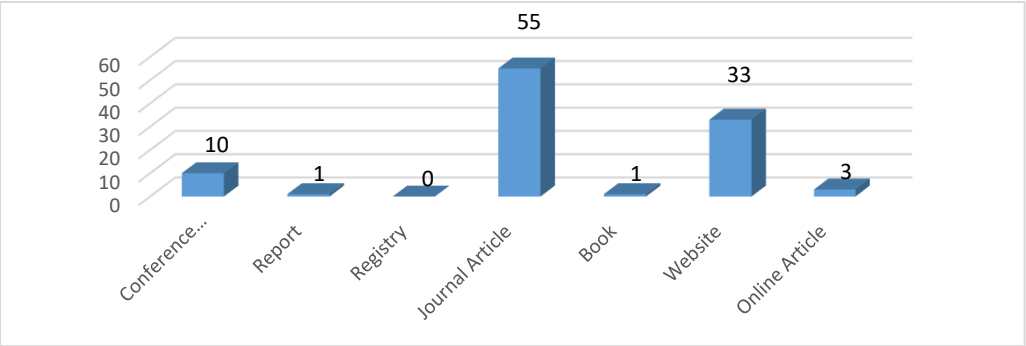


**Figure 2: Distribution of used literature review**

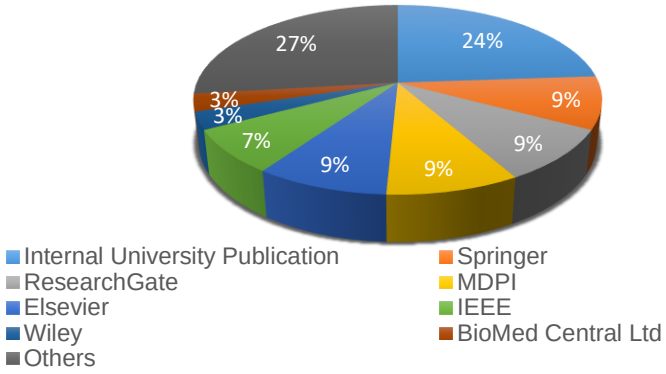


**Figure 3: Distribution of database and registry used**

There are eighteen other databases and registries in which only one article or publication is used in this study. This includes SCOPUS, Arxiv, ERIC, Emerald, Kamla-Raj Publisher, IOP Publishing, Hindawi Limited, Apress Media LLC, Indonesian Journal of Educational Research and Technology, Psicothema, Journal of Computer Science and Information Technology, Oxford Academic, European American Journal, Central and Eastern European Online Library, BMJ, International Journal of Advanced Computer Science and Applications (IJACSA), Frontiers and Modestum LTD. Besides, the highest number of articles used were from the university's internal publication, which contains sixteen articles. The distribution of articles found based on the topic is shown in Figure 4.

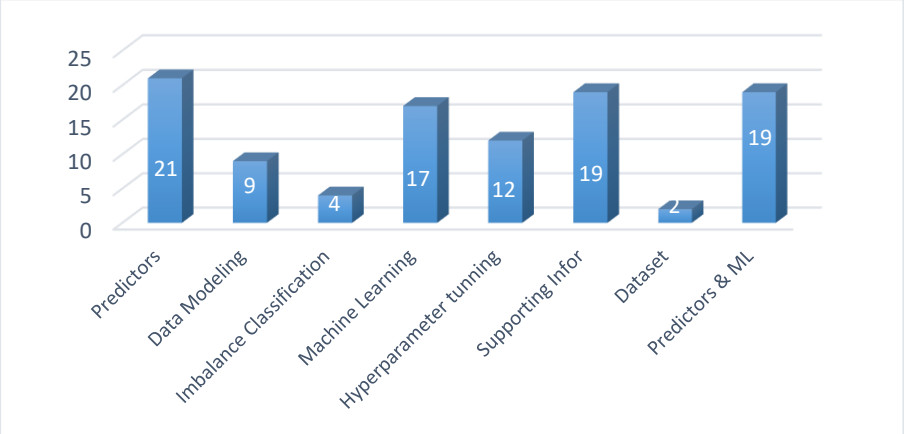


**Figure 4: Distribution of found articles based on topic**

Most articles found were on the predictors of student academic performance, this is followed by articles which include predictors and machine learning and supporting information. Supporting information is the articles which contain the information which explains the needs of the research,

and explanation of the technology used in this study. The third and fourth-highest articles used are on machine learning and hyperparameter tuning.

## 3. Predictors for academic performance

Education transmits values and accumulated knowledge of society (Naka et al., 2003). Education is the process of teaching or learning, especially in school or college. A student is a person who is learning at a school (Cambridge University, 2023). During education, the student is the main character supported by family and teachers. An educational institution can also be called a school, college, or university. The student's academic performance measures their achievement (Ballotpedia, 2023). In this study, the research is entirely focused on academic performance obtained via a standardised test that evaluates the progress and standards of the students. A secret relationship exists between students, their teachers, their families, and the school, causing students to perform differently academically. In this study, the relationship between each factor will be unfolded. All factors (predictors of academic performance) that affect students' academic performance are categorised into four main groups: student-related, family/parents-related, teacher-related, and school-related. Researchers named the parameters differently based on the resources they referred to. All the factors have been put into sub-categories for easy analysis and comparison.

### 3.1 Student-related predictors

The student is the main character in the research. Most listed factors affect the student's academic performance from the "Student" category. Table 2 is the list of previous studies done on the factors affecting students' academic performance related to students.

**Table 2: List of the previous study which student-related predictors**

| Factor | Description | Previous Research |
|---|---|---|
| ***Characteristic*** | Age, gender, nationality, place of birth, height, disability status, learning deficit, severe previous health issues, mental health, low emotional resilience, fear, stress, relationship with others, social activity status, current discipline division. | Abdullah et al., 2022; Agnafors et al., 2021; Aman et al., 2019; Arashpour et al., 2023; Arora & Singh, 2017; Briones et al., 2021; Dabhade et al., 2021; Feng, 2019; Inusah et al., 2022; Jalota & Agrawal, 2019; Lau et al., 2019; Melo et al., 2022; Niu et al., 2022; Rajalaxmi et al., 2019; Tadese et al., 2022; Widyaningsih et al., 2019; Yakubu & Abubakar, 2022; Zhang et al., 2021 |
| ***Attitude*** | Towards learning, towards going to school, towards study, towards attending class, towards the assignment, towards a subject, discipline behaviour, distraction factor, absences rate, raising a hand in class, attempts count in passing a subject, attitude in class, opening study resources, view announcement, discussion group. | Aman et al., 2019; Arashpour et al., 2023; Briones et al., 2021; Jalota & Agrawal, 2019; Z. Li & Qiu, 2018; Niu et al., 2022; Olufemioladebinu et al., 2018; Owusu-Boadu et al., 2021; Sokkhey & Okazaki, 2020; Widyaningsih et al., 2019; Zulfiker et al., 2020 |
| ***Habit*** | Sleeping time, leisure time, smoking habit, time spent on computer games, addiction to online platforms, often using WhatsApp. | Arora & Singh, 2017; Dabhade et al., 2021; Olufemioladebinu et al., 2018; Owusu-Boadu et al., 2021; Tadese et al., 2022; Widyaningsih et al., 2019 |
| ***Interest*** | Perform presentations, fitness activities, and sports. | Dabhade et al., 2021; Zulfiker et al., 2020 |
| ***Learning Experience*** | Previous discipline division, previous discipline courses, location of primary school, the medium used in primary school, type of primary school, first education starts on age. | Bayirli et al., 2023; Dabhade et al., 2021 |

| | | |
|---|---|---|
| ***Pass Academic Achievement*** | GPA, Credit awarded, entry exam result, course score, semester score, prior qualification, average high school grade points, mid-term exam. | Arashpour et al., 2023; Feng, 2019; Hamoud et al., 2018; Huynh-Cam et al., 2022; Lau et al., 2019; Z. Li & Qiu, 2018; Yağcı, 2022; Yakubu & Abubakar, 2022; Yohannes & Ahmed, 2018 |
| ***Satisfaction*** | Towards school | Jalota & Agrawal, 2019; Z. Li & Qiu, 2018 |
| ***Others*** | English fluency, illegal activities, teenage pregnancy and motherhood, distractive factor. | Arora & Singh, 2017; Dabhade et al., 2021; Melo et al., 2022 |

Based on Table 2, out of thirty-eight literature reviews of predictors for student academic performance that have gone through, thirty literature reviews focus on student-related factors. The sub-categories related to students are *Characteristics*, *Attitude*, *Habit*, *Interest*, *Learning experience*, *Past academic achievement*, *Awareness*, *Knowledge*, and *Satisfaction*. *Characteristics* of the students are features that show a student's personality. These include age, gender, nationality, place of birth, height, disability status, severe previous health issues, mental health status, low emotional resilience, fear, stress, relationship with others, social activity status, and current discipline division. Among all the research, there are seven which put their focus on age as one of the factors (Aman et al., 2019; Feng, 2019; Inusah et al., 2022; Lau et al., 2019; Niu et al., 2022; Rajalaxmi et al., 2019; Tadese et al., 2022). On the other hand, gender from different ages has been highlighted in seven research (Agnafors et al., 2021; Arashpour et al., 2023; Bayirli et al., 2023; Widyaningsih et al., 2019; Yakubu & Abubakar, 2022; Zhang et al., 2021). They found that girls are more likely to make themselves more eligible for higher education (Agnafors et al., 2021). A student's nationality or place of birth can be a factor that affects a student's academic performance. This has been highlighted in Zhang et al. (2021), Niu et al. (2022), and Jalota & Agrawal's (2019) research. Physical health-related factors like disabled status and severe health issues that happened to students before and mental health statuses like fear, stress, well-being, and emotional management, which include low emotional resilience, can be part of the predictor (Abdullah et al., 2022; Agnafors et al., 2021; Arashpour et al., 2023; Dabhade et al., 2021; Melo et al., 2022). On the other hand, the social skills of a student-like relationship between people around the students (parents, siblings, teachers, and classmates) have been put as one of the factors (Arora & Singh, 2017; Feng, 2019) The course that the student currently registers for is part of the student's characteristics and has been considered as one of the factors (Niu et al., 2022). In line with this, the number of elective credits as a predictor is found in research done by Huuyng-Cam et al. (2022). This happened in Jalota & Agrawal (2019) and Aman et al. (2019) research, which takes into consideration the current student's educational stage, grade level, and section.

Another sub-category that relates to students is *Attitude*. Twelve research found that the attitude of students towards learning, subjects, assignments, and class, like listing important notes, concentrating in class, playing after finishing homework, checking their answers after finishing homework, student laziness to study, studying only when they like, student copies the assignment of a friend, putting more effort to learn complex concepts, attentive during class, number of times student raised hands in class, attendance to school, number of attempts in passing a subject, submission of assignment, opening online resources, having discussion group with classmate view announcement can significantly affect the student academic performance (Arashpour et al., 2023; Arora & Singh, 2017; Hamoud et al., 2018; Jalota & Agrawal, 2019; Z. Li & Qiu, 2018; Morilla et al., 2020; Niu et al., 2022; Olufemioladebinu et al., 2018; Owusu-Boadu et al., 2021; Sokkhey & Okazaki, 2020; Widyaningsih et al., 2019; Zulfiker et al., 2020).

**Student *Habit*** is another sub-category that is related to students. A student's habit can be good or bad. Good habits include studying regularly, frequency of going to the school library, time arrangement to study at home, sleeping time, and leisure time, which is in good plan (Arora & Singh, 2017; Dabhade et al., 2021; Olufemioladebinu et al., 2018; Owusu-Boadu et al., 2021; Widyaningsih et al., 2019). On the other hand, Dabhade et al. (2021) and Tadese et al. (2022) also mentioned other bad habits like time spent on computer games, addiction to online platforms, frequent use of WhatsApp, and smoking habits.

Another sub-category that relates to students is student *Interest*. Interest draws the student's attention, and they are willing to spend time on it when they have free time (Dictionary.com, 2023). Interests like performing presentations, fitness activities, and sports have been highlighted in the research of Dabhade et al. (2021) and Zulfiker et al. (2020). Meanwhile, a student's interest in a subject like mathematics can also be a factor (Sokkhey & Okazaki, 2020).

***Learning Experience*** is another sub-category that is related to students. The learning experience is significant to consider as a factor affecting students' academic performance. The learning experience is the student's knowledge foundation before proceeding. Previous discipline division and course, primary school location, language used in primary school, type of primary school, and first education starts on age are the factors categorised under learning experience (Bayirli et al., 2023; Dabhade et al., 2021).

Besides, ten pieces of research also mentioned that students *Pass academic achievement*, for example, GPA and credits earned factors that will help to predict the student academic performance (Briones et al., 2021; Hamoud et al., 2018; Lau et al., 2019; Z. Li & Qiu, 2018; Morilla et al., 2020; Yağcı, 2022; Yakubu & Abubakar, 2022; Yohannes & Ahmed, 2018; Zulfiker et al., 2020). All this has been categorised in the past academic achievement sub-category related to students. The *Satisfaction* sub-category related to the students is only mentioned in Li & Qiu (2018) and Jalota & Agrawal (2019). The satisfaction of students towards school, class advisor, and teacher can be a factor affecting student academic performance (Jalota & Agrawal, 2019; Z. Li & Qiu, 2018).

There are *Others* sub-category groups with those factors that cannot be categorised into the abovementioned sub-category. In this category, Dabhade et al. (2021) consider English fluency as one of the factors affecting the student's academic performance. Besides, Melo et al. (2022) also highlighted illegal activities done by students, teenage pregnancy, and motherhood as factors affecting student academic performance (Melo et al., 2022). Arora & Singh (2017) highlighted those distractions like mobile phones, the internet, and peer pressure are the second highest factors affecting students' academic performance.

Looking at over thirty related works, they try to discover possible factors, but some loopholes are yet to be considered. In this study, there will be one more sub-category: *Past experiences* like having the student being bullied, *Awareness* like own body image, the meaning of life, and global issues, which should be considered as factors affecting students' academic performance. Besides, more factors can be categorised as *Characteristics* of students not covered by other researchers, such as body mass of the student, immigration status, spoken language at home, well-being of students, cognitive flexibility, competitiveness, credibility, resilience, confidence level of handling the financial matter and confident level of taking financial matter using digital device and respect people from other cultural. On the other hand, factors that can be categorised into *Attitudes* which do not include in further research are attitudes towards immigrants, attitudes towards future careers, attitudes towards social interaction, and students' attitudes that will hinder their learning in school, which can be put into the study as part of the factors. Besides, there are some missing *Interests*, like learning about other cultures, and interest in ICT was not mentioned in the previous study. Good *Habits* like enjoying reading are not mentioned as one of the factors either. On *Learning experience*, the number of changes in educational organisation are not noted as part of the predictors. Therefore, this research is vital to fill the gap in the previous study.

### 3.2 Family/parents-related predictors

Parents and family members are the people students deal with when they are at home. Table 3 lists previous studies on the factors/predictors affecting students' academic performance related to family/ parents.

**Table 3: List of previous studies which are family/parents-related predictors**

| Factor | Description | Previous Research |
|---|---|---|
| ***Family member at home*** | Number of siblings, number of people residing in the household | Dabhade et al., 2021; García & Skrita, 2019 |
| ***Socioeconomic*** | Income, family social status, family socioeconomic, parents' average income, parents' income, parents' occupation, parents' educational background, place of living (stratum of household) | Bayirli et al., 2023; Briones et al., 2021; Dabhade et al., 2021; García & Skrita, 2019; Huynh-Cam et al., 2022; Lau et al., 2019; Z. Li & Qiu, 2018; Olufemioladebinu et al., 2018; Ozcan, 2021; Sokkhey & Okazaki, 2020 |
| ***Attitude*** | Parenting style, parental education participation, parents, attitude towards students' study, parents' attention to student attitude, parent time and money student, parents' response to student attitude, parent's encouragement, parental support, motivation from a family member | Abdullah et al., 2022; Bayirli et al., 2023; Briones et al., 2021; Dabhade et al., 2021; Z. Li & Qiu, 2018; Ozcan, 2021; Sokkhey & Okazaki, 2020 |
| ***Atmosphere of environment*** | Region of student staying, residence area, distance from home to school, learning environment at home, | Bayirli et al., 2023; Z. Li & Qiu, 2018; Sokkhey & Okazaki, 2020 |
| ***Resources at home*** | Predominant material, cultural resources, computers for schoolwork, books for study at home, electric appliances at home, | Bayirli et al., 2023; García & Skrita, 2019 |

Only eleven of the literature reviews contain predictors related to family members or parents. This group of predictors has further divided into five sub-categories. These sub-categories included *Family members at home*, *Socioeconomic*, *Attitude*, *Atmosphere of staying environment*, and *Resources at home*. In the aspects of *Family members at home*, only Garcia & Skrita (2019) used the number of siblings and the number of people residing in the same house to predict students' academic performance. On the other hand, Dabhade et al. (2021) listed students staying with family members at home as one of the factors affecting the student's academic performance.

The most common sub-categories related to family is *Socioeconomic*. Socioeconomic status (SES) is an individual's position in a society's hierarchy, which indicates an individual's contribution to the society and economy. This includes income, education level, type of occupation, place of residence, and original ethnic and religious background (American Psychological Association (APA), 2022). There are ten literature reviews in their study, including this sub-category. Li & Qiu (2018), Sokkhey & Okazaki (2020), and Dabhade et al. (2021) put socioeconomic in their study. Li & Qiu (2018) highlighted that out of all the factors in their research, family socioeconomic status has the most impact on students' academic performance. Besides these two research studies, they look at different aspects, like parents' income. Briones et al. (2021), Huyng-Cam et al. (2022), and Gracia & Skrita (2019) have included parents' income as a predictor. Besides, parents' occupations are also included as a predictor in Huyng-Cam et al.'s (2022) and Lau et al.'s (2019) study. Parent's educational background is highlighted in Olufemioladebinu et al. (2018), Garcia & Skrita (2019), Bayirli et al. (2023), Li & Qiu (2018) and Huyng-Cam et al. (2022). Among these four research studies, Li & Qiu (2018) have proven that parents' educational background is the third predictor of the predictors that will impact students' academic performance. There is a rare factor in which type of houses the students stay, for example, low-cost flats, condominiums, terraced houses, semi-detached houses, and bungalows as a predictor in the study (García & Skrita, 2019).

Besides, the *Attitude* of family members is another sub-category that falls in the group of family-related predictors. The mentioned attitude is not only from parents to students but also from other family members like siblings. Briones et al. (2021) included parenting styles like supportive, strict, and democratic styles in the student study. The result was that parents with a supportive style would have children with good academic performance (Briones et al., 2021). Li & Qiu's (2018) research

found that parental education participation is the fourth predictor of students' academic performance. Bayirli et al. (2023) listed parents' attitudes toward student study as one of the predictors in their research. Parents' attention to student attitude, parents' response to student attitude, and parents' encouragement to their children are mentioned in Sokkhey & Okazaki's (2020) study as a few of the predictors showing the parents' attitude toward the student. The willingness of parents to spend time and money on students is also considered one of the aspects of parents' attitudes towards their children (Sokkhey & Okazaki, 2020). Educational support from family can be one of the factors that affect a student's academic performance (Ozcan, 2021). Motivation from family has been listed as one of the predictors in both Dabhade et al.'s (2021) and Ozcan's (2021) studies.

Another sub-category is the *Atmosphere of the environment*, including the region of the student's staying, residence area, distance from home to school, and learning environment at home. Li & Qiu (2018) and Garcia & Skrita (2019) have separately mentioned the region where students stay and the residence area, which are related to the atmosphere of the student's housing area. Sokkhey & Okazaki (2020) is the only research that put the distance between home and school as one of the predictors. The learning environment at home has been listed as one of the predictors in Bayirli et al.'s (2023) and Sokkhey & Okazaki's (2020) studies.

*Resources at home* is another sub-category, which is family-related predictors. Garcia & Skrita (2019) listed two main groups of resources at home: predominant materials like washers, microwaves, ovens, and cars, and cultural resources like the number of books at home, DVD players, Internet access, television, and computers in their research. On the other hand, Bayirli et al. (2023) mentioned computers and electric appliances at home as among the factors affecting students' academic performances.

### 3.3 Teacher-related predictors

Table 4 lists previous studies on the factors affecting students' academic performance related to teachers.

**Table 4: List of previous studies which are teacher-related predictors.**

| Factor | Description | Previous Research |
|---|---|---|
| ***Characteristic*** | Gender | Inusah et al., 2022 |
| ***Educational background*** | Qualification of teacher, education level, professional development course, average hour of professional development | Weaver, 2019 |
| ***Working experience*** | Teaching experience | Weaver, 2019 |
| ***Attitude*** | Absence from class, commitment to duty, as a role model to student, encouragement to students, motivation to students, connection with students, adequate help to students, guidance to students, emotional support, academic support | Abdullah et al., 2022; Isaac, 2019; Ozcan, 2021; Rose, 2020; Sokkhey & Okazaki, 2020 |
| ***Skillset*** | Communication skill, subject mastery, effectiveness in conducting class, teaching method, planning for class, ability to teach in class, use of handouts in class, teacher administration of student assignment, teacher-student interaction, teacher competency, questioning behaviour | Briones et al., 2021; Isaac, 2019; Olagbaju, 2020; Olufemioladebinu et al., 2018; Ozcan, 2021; Sokkhey & Okazaki, 2020 |
| ***Others*** | Lack of commitment, burnout | Rose, 2020 |

There is a total of ten literature reviews that mention teacher-related predictors. The teacher-related predictors have been further divided into sub-categories of *Characteristics*, *Educational background*, *Working experience*, *Attitude*, *Skillset*, and *Others*. Only one previous research included the *Characteristics* of an educator, which is the availability of educators by gender in their study (Inusah

et al., 2022). The *Educational background* sub-category includes the educator's qualification and the on-the-job training they undergo during their service in an educational organisation. The only research paper from Weaver (2019) studies the relationship between educators' educational backgrounds and student performance in the American College Test (ACT). The study has considered the teacher's qualification level, the type of professional development course, and average hours of professional development as the predictors. However, the research found no significant relationship between an educator's educational background and student performance in ACT (Weaver, 2019).

For the *Working experience* sub-category, the only literature review that includes this sub-category in the research is Weaver (2019). The study also found that an educator's years of service in an educational organisation are not significantly related to the student's ACT performance (Weaver, 2019). Another sub-category is *Attitude*, which includes the educator's attitude towards work and students. Predictors that indicate the educator's attitude towards work have been mentioned in both Sokkhey & Okazaki (2020) and Rose (2020) on an educator's absence from work. Besides, Isaac (2019) also puts educators' commitment to their duty as part of the predictor, and Ozcan (2021) puts educators as students' role models as one of the predictors in the research. On the other hand, educators give encouragement and motivation to students in Sokkhey & Okazaki (2020) and Ozcan (2021) separately. Sokkhey & Okazaki (2020) also include connections with students; educators provide adequate help to students as predictors in their study. Abdullah et al. (2020) highlighted emotional support and academic support from educators as predictors in their research.

*Skillset*s are another sub-category of predictors that are teacher-related. This sub-category contains both soft skills and professional teaching skills of the educator. Soft skills like communication skills are mentioned in Olufemioladebinu et al.'s (2018), Isaac's (2019), and Ozcan's (2021) research. Besides, professional teaching skills like subject mastery are highlighted by Olufemioladebinu et al. (2018), Sokkhey & Okazaki (2020), and Olagbaju (2020). Based on results from Olufemioladebinu et al. (2018), a group of students with more than half had average results. They commented that most educators have good communication skills and can master the subject. In the same research, students also commented that educators' class management skills are essential to them (Olufemioladebinu et al., 2018). Sokkhey & Okazaki (2020) also highlighted that an educator's ability to handle class, involvement in preparing the class, and use of handouts during class are part of the predictors. Besides, Isaac (2019) also highlighted that educators' administration of students' assignments is one of the factors. Questioning behaviour is only mentioned in Olagbaju (2020) as one of the predictors. The other sub-category needs more commitment and burnout, indicating the workload of educators in the educational organisation. Both predictors were found in Rose's (2020) research report.

### 3.4 School-related predictors

Table 5 lists previous studies on the factors affecting students' academic performance related to school.

**Table 5: List of previous studies which are school-related predictors**

| Factor | Description | Previous Research |
|---|---|---|
| ***School Environment*** | Safety around the school area, location of the school, distractive factors around the school | Agnafors et al., 2021; Inusah et al., 2022; Ozcan, 2021 |
| ***School Facility*** | Internet stability, libraries, water supply, electricity, playground, staff office, hostels, facility for students, facility for teacher, condition of school facility | Briones et al., 2021; Inusah et al., 2022; Olufemioladebinu et al., 2018; Ozcan, 2021 |
| ***Management*** | Classroom arrangement, class arrangement, teacher workload, teacher salary, teacher incentive, school management effective, school administration | Bayirli et al., 2023; Isaac, 2019; Olufemioladebinu et al., 2018; Ozcan, 2021; Rose, 2020; Sokkhey & Okazaki, 2020 |
| ***Others*** | Study resources in school | Inusah et al., 2022 |

The previous research that included school-related predictors in the study divided the predictors into four sub-categories. This includes *School environment*, *School facility*, *Management*, and *Others*. In the sub-category of the *School environment*, Agnafors et al. (2021) mentioned that the safety of the school environment will cause students to have negative experiences, which will lead to students not having good academic performance. On the other hand, Inusah et al. (2022) put the school's location as part of the predictor during the research. Ozcan's (2021) study considered distractive factors around the school as one of the predictors.

In the sub-category of *School facility* for some of the research, the focus is on the facility as a whole. Some other research focuses on the facility to the user group, like the facility for students and teachers. Briones et al. (2021) stated that adequate internet access significantly impacts student's academic performance. Olufemioladebinu et al. (2018) listed identical facilities like libraries, water supply, electricity, playgrounds, staff offices, and hostels as predictors. Inusah et al. (2022) grouped the school facilities into three main groups: school facilities for educators, students, and study resources in school. Ozcan (2021) highlighted the condition of school facilities as a predictor in the research.

*Management* is another sub-category of school-related predictors. School management is all the arrangements based on school resources and policy. This includes classroom arrangement, class arrangement, arrangement of resources, teacher workload, teacher's salary, teacher's incentive, and arrangement of content for study. Olufemioladebinu et al. (2018), Sokkhey & Okazaki (2020), and Isaac (20119) have mentioned classroom arrangement as one of the predictors. Classroom management must depend on the available classrooms, educators, and the number of students in the school. However, as mentioned in Bayirli et al. (2023) and Sokkhey & Okazaki (2020), class management is the arrangement of different subjects in a week as part of the predictors. Both Isaac (2019) and Rose (2020) mentioned that the workload of educators is part of school management and arrangement of work for educators, including teaching and administration work. Besides, Isaac (2019) also mentioned the educator's salary and educator's incentive, which might affect the educator's work contribution and indirectly impact the student's academic performance. There are two notable points of view that Ozcan (2021) includes in the study of the effectiveness of school management and school administration. In *Others* sub-categories, study resources like English, maths, science, and social studies textbooks in school have been mentioned in Inusah et al. (2022) as one of the predictors.

### 3.5 Literature review with predictors from four categories

Based on all the literature reviews done, three of the literature reviews contain predictors from all four categories. Table 6 lists the sub-categories available in the research and the predictor most significantly affecting the student's academic performance. Based on the summary Table 6, although all three of the stated literature reviews have predictors from four categories, each includes different sub-categories. Besides, the detailed variables in the sub-categories are different too. For example, three of the literature reviews have sub-categories of *Socioeconomics*, but Briones et al. (2021) only refer to parents' income. In comparison, Sokkhey & Okazaki (2020) refer to parents' educational background and occupation status. Olufemioladebinu et al. (2018) refer to parents' educational backgrounds. Besides, the objective and the scope of research are different, and the outcome of the study can also be different. Data collected during Briones et al. (2021) research are from Grade 11 students at SKSU-Laboratory High School. The research concludes that students who lack motivation have the highest impact on the student's academic performance.

**Table 6: List of literature review which contains predictors from four categories.**

| Reference | Student-Related Predictor | Family-Related/ Parents-Related Predictor | Teacher-Related Predictor | School-Related Predictor |
|---|---|---|---|---|
| Briones et al., 2021 | Characteristic, Attitude | Socioeconomic, Attitude | Skillset | School Facility |

| | | | | |
|---|---|---|---|---|
| Sokkhey & Okazaki, 2020 | Attitude | Socioeconomic, Attitude, Atmosphere of environment | Attitude, Skillset | Management |
| Olufemioladebinu et al., 2018 | Habit | Socioeconomic | Skillset | Management, School Facility |

Sokkhey & Okazaki's (2020) research aims to prepare a hybrid approach using principal component analysis (PCA) to improve students' academic performance prediction. The dataset used in this research is collected from twenty-two high schools in Cambodia. The study did not highlight any main factors affecting the student's academic performance. Meanwhile, Olufemioladebinu et al. (2018) aim to assess the seven selected predictors affecting student academic performance. The dataset used in this research includes six selected education colleges in Southwest Nigeria. The study did not highlight that any of the used predictors carries the highest weightage among the predictors.

## 4. DATA MODELING

Data modelling is a process that identifies the predictors and all the relationships between them. This process structurally places all the predictors to enable the model to achieve the shortest processing time with the highest accuracy. There are three levels of data abstraction:

1. Conceptual level: This level defines the high-level dataset and relationship of the collected. This level of abstraction is presented on a diagram as a visual presentation.
2. Logical level: This level defines the relationship and constraints between datasets. This level of abstraction is presented in modelling languages like SQL or ER Diagram.
3. Physical Level: This level defines how each predictor is stored and includes the data type, indexing and other technical designs.

With proper data modelling, the dataset can better understand the relationship between all the datasets in this study. This is to build a new dataset for the machine learning algorithm. Besides, the data modelling process will help remove inconsistent and error data. This will help to prevent wrong data from being used to train the model and predict the result. Improving collaboration between datasets is a benefit of doing data modelling. This ensures that all the datasets linked together will allow the model to produce higher accuracy predictive results. With a clear and consistent dataset, the algorithm model can reduce the processing power and use a shorter time to make predictions (Simplilearn, 2023).

There are two ways that all the datasets can be linked together. It can be either using star schema or snowflake schema. Star schema (Figure 5) has a fact table and a few dimension tables. The fact table is in the centre, which stores the primary data, for example, all the predictors in the dataset. The dimension table holds the supporting information. In the final data model, there will only be one fact table and more than one dimension table. The dimension table in the star schema should not contain any foreign key as the data do not refer to any other table. Using the star schema denormalises data in the dataset to avoid duplication in the dimension table (Tech Target Contributor, 2024).

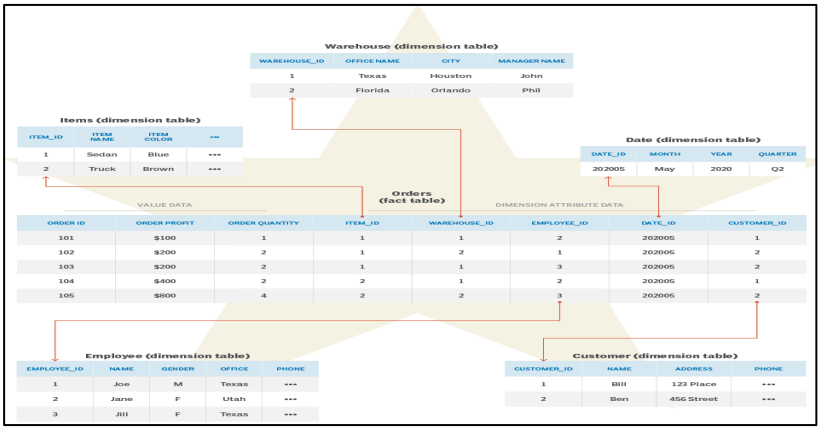


**Figure 5: Sample Star Schema (Tech Target Contributor, 2024).**

Snowflake data schema (Figure 6) is similar to star schema in that they have only one fact table and multiple layer dimension table. However, the dimension table in the snowflake data schema has a foreign key that refers to another dimension table. A dimension table in Snowflake can be branched to another table. Snowflake data schema is more normal than star schema (Tech Target Contributor, 2024).

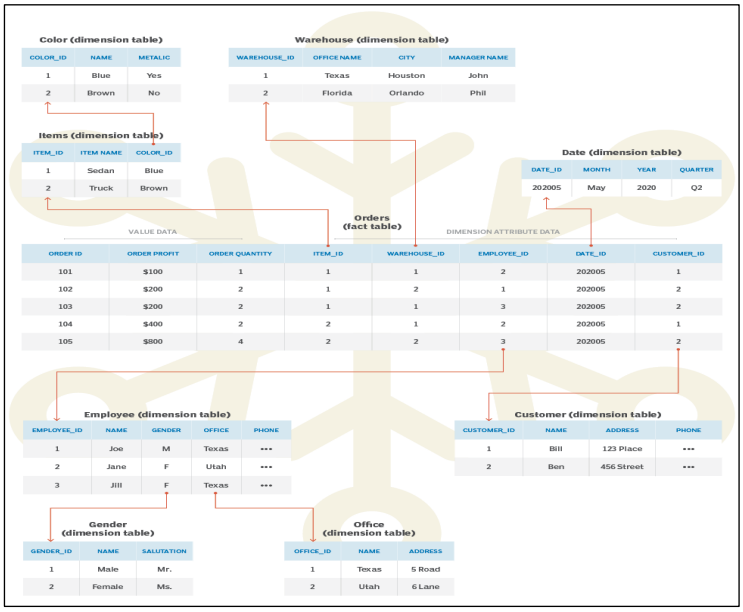


**Figure 6: Sample Snowflake Schema (Tech Target Contributor, 2024)**

Three previous studies highlighted that star schema could speed up the analytic work to assist decision-making as presented in Table 7 (Amin et al., 2021; Hoi et al., 2023; Nazir et al., 2022). Nazir et al. (2022) mentioned that star schema can minimise data redundancy in dimension tables. Besides, star schema is also an optimised model that is the best performance for presenting complex information (Mandala et al., 2021). Suharso et al. (2021) mentioned that star schema is suitable for Online Analytical Processing (OLAP). There are ready-to-use open-source plugins from Python Library, which can integrate and extract information from star schema.

**Table 7: List of literature review for star schema**

| Reference | Field | Data Size |
|---|---|---|
| Hoi et al., 2023; Mandala et al., 2021; Nazir et al., 2022 | Enterprise | Contain structured, semi-structured and unstructured data |

| | | |
|---|---|---|
| Amin et al., 2021; Suharso et al., 2021 | Education | An increasing number of records in existing databases cannot support rapid decision-making. |
| Morassi Sasso et al., 2021 | Medical | 8 million patients, 4 billion records |

Star schema is a good choice for complex and large datasets as mentioned in the previous research. However, the datasets have never implemented this in predicting students' academic performance. Star schema is needed for this study as the dataset combines three different datasets containing around two million records. This number is lower than four billion entries in the hospital patient record (Morassi Sasso et al., 2021). This confirms that star schema can perform well in this study and provides extra capacity for future data to be loaded into the schema.

## 5. IMBALANCE CLASSIFICATION

Imbalance classification means data distribution in different classes is skewed (Brownlee, 2020). The classification can have a slight unevenness, like having four against another group of six, which is known as a slight imbalance. Severe imbalance is highly uneven data distribution, like one against one hundred and some against one thousand (Krawczyk, 2016). Slight imbalance classification will generally be treated as usual. A severe imbalance will need more attention as that will cause bias in the predicted result. The class that contains the most data is known as the majority class, whereas the class with the most minor data is known as the minority class. When the model is trained with the majority class, more patterns and sample data will be learned by the model. Therefore, the predicted result will have a higher possibility of falling into the majority class. On the other hand, the minority class contains less data. During the learning process of the model, the model will have less sample data and patterns to learn. This caused the prediction result to have a lesser chance of falling into the minority class. In the event of bias, problems like fraud detection, claim prediction, default prediction, churn prediction, spam detection, anomaly detection, outlier detection, intrusion detection, and conversion prediction will happen. Therefore, processes must be improved to prevent bias in the prediction result (Brownlee, 2020). Seven techniques can be used to handle an imbalance dataset:

### Evaluation metric

There are a few different evaluations to select the predictor to train the model accordingly. Most of the evaluation metric is to check how relevant the factor is to the expected predicted result. The selection of evaluation type depends on the data in predicting the result. Commonly used evaluations are precision, recall and F1 score. One can also use the Matthews correlation coefficient (MCC) to check the coefficient between predictors and predicted results for predicted results in binary classification. The ROC curve (AUC) is another type of evaluation to determine the relation between the true-positive rate and the false-positive rate (Wu & Radewagen, 2022).

### Resample training dataset

There are two approaches, under-sampling and over-sampling, which can be done to make the imbalanced dataset a balanced dataset. Under-sampling is to be done by reducing the size of the majority class and keeping all data in the minority class. In this way, a new balanced dataset will be generated. However, this method can only be applied when the dataset has sufficient data. On the other hand, over-sampling increases the data in minority classes. New data can be generated using repetition or bootstrapping (Wu & Radewagen, 2022).

### K-fold cross-validation

Applying K-fold cross-validation to an imbalanced dataset will cause overfitting of the majority class. Therefore, applying K-fold cross-validation on imbalanced datasets requires special handling. The dataset is stratified so that the primary data contains the details of the minority class. K-fold cross-validation needs to be implemented with over-sampling, but over-sampling is only applied to the training set in each cycle of cross-validation evaluation (Akiode, 2020).

### Ensemble different resampled dataset

Using all the samples from the minority class and different random sampling from the majority class to implement in different models. This approach is mainly horizontally scaled for the majority class to train and run the model. This method is an ensemble model which can produce a generalised prediction result (Wu & Radewagen, 2022).

### Resample with different ratio

This method is the same as the one discussed in the previous section, but the ratio of minority and majority will be using different combinations depending on data and models. This can be 1:3 (minority class: majority class) or 2:1. This will help change the weightage of the class.

### Cluster the majority class

The majority class is clustered into different groups. The cluster centroid is the average feature vector for all the data in the clustered majority class. The data farthest from the cluster centroid can be removed as part of the data as that indicates less importance than the data nearest to the cluster centroid (Chakrabarty, 2018).

### Design own model

This method focuses on the selected model. A model like extra gradient boost has a feature which handles imbalanced datasets by resampling data during model preparation. Besides, weightage can be implemented for the minority class to generalise the minority class naturally. With this method, bias will not happen in the result of the prediction (Wu & Radewagen, 2022).

Table 8 lists a literature review on handling imbalanced datasets.

**Table 8: List of Literature reviews found on how to handle the imbalanced dataset.**

| Reference | Handling imbalanced dataset |
|---|---|
| Injadat et al., 2020; Kaensar & Wongnin, 2023; Melo et al., 2022 | Resampling training dataset: SMOTE |
| Bayirli et al., 2023 | Resampling training dataset: Randomly downsampling |
| Niu et al., 2022 | Resampling training dataset: SMOTE<br>Evaluation metric: finding out the correlation of attribute<br>Design own model: Radial Basis Function and Naïve Bayes |

In the previous research from the same field on predicting student performance or results, four studies have the same problem of imbalance classification in the dataset. The dataset in Melo et al. (2022) is from nineteen universities within a state in Brazil. During the data preparation stage, an imbalance classification was found. The researchers have implemented the Synthetic Minority Over-sampling Technique (SMOTE) in the minority class to fill the gap between the minority and majority classes (Melo et al., 2022). Injadat et al. (2020) noticed that the dataset only has fifty-two data that can be used. However, they decided not to implement SMOTE in the dataset as the researchers would want to maintain the real-life data without overfitting it. On the other hand, in the discussion, the researcher suggested that SMOTE should be implemented in the minority class (weak students) to balance the data with the majority class (Injadat et al., 2020). Likewise, Bayirli et al. (2023) also found an imbalance distribution of classified Pausible Value for Mathematics in the PISA dataset. Therefore, randomly down-sampling on the majority class has been implemented in the dataset. The main objective of Niu et al. (2022) is to handle the imbalanced dataset, causing biased results in the prediction. This research uses two datasets. One is that the dataset is from a university in Beijing, China. Another dataset is a public dataset, which is Kalboard 360. In this research, three methods are being implemented to handle the imbalanced dataset. During the preparation of data, SMOTE on the minority class. Besides, the researchers also carry out correlation attribution evaluation to find the most relevant factor. Lastly, the result is gained through two models: Radial Basis Function and Naïve Bayes (Niu et al., 2022).

In the previous study, four researchers used SMOTE to upsize the minority class. Bayirli et al. (2023), who are working on the PISA dataset, focus only on Plausible Values for mathematics and use a down-sizing method to reduce the amount of data in the majority class to get a balanced dataset during the preparation of the dataset. The researchers have not fully utilised all three datasets in PISA. They are only using student datasets. The Plausible Value for mathematics is distributed into two classes only.

## 6. MACHINE LEARNING-BASED ACADEMIC PERFORMANCE PREDICTION RESEARCH

Machine learning is a branch of artificial intelligence (AI) and computer science that focuses on using data and algorithms to imitate how humans learn, gradually improving the accuracy of the model (IBM, 2023). Computer programs adapt and evolve based on the data they process to produce predetermined outcomes. They are essentially mathematical models that "learn" from fed data (training data) (ARM, 2023). Machine learning is divided into the decision process, error function, and model optimisation. Firstly, a machine learning algorithm will make a prediction or classification. In other words, based on the input data, the algorithm will produce an estimated pattern in the data. Then, a comparison will be conducted to assess the model's accuracy. Once the model with the highest accuracy is found, the model will then be adjusted to fit better to the data points in the training set. This process will be repeated until it reaches a threshold of accuracy (IBM, 2023).

There are three types of machine learning algorithms: Supervised Learning Algorithms, Unsupervised Learning Algorithms, and Reinforcement Learning Algorithms. Supervised learning algorithms consist of a dependent variable expected to be predicted from a given set of independent variables. Supervised learning algorithms are trained with labelled input and output data (Shee, 2022). All the supervised learning algorithms can be further divided into two groups based on the type of prediction result. If the prediction result type is in quantitative data like the amount of income, a person's body weight, or height, then this is regression. The algorithms that can predict regression results are linear regression and logistic regression (Shee, 2022). On the other hand, if the prediction separates the data into discrete values or multiple classes, then it is called classification. Sample classification algorithms are binary classification, multi-class classification, and decision tree. Unsupervised Learning Algorithms are being trained with raw and unlabeled data (Shee, 2022). Examples of unsupervised learning algorithms are Apriori algorithms and K-means clustering. Reinforcement Learning Algorithms are used to train machines to make specific decisions. The machine will be exposed to an environment where it continuously uses trial and error to prepare itself. The machine learns from experience and captures the best possible knowledge to make accurate business decisions. An example of Reinforcement learning is the Markov Decision Process (Sunil Ray, 2023). In this research, the student results are categorised into different categories. Therefore, this research will use all of the algorithms under supervision and classification type of algorithm. Following are some of the commonly used algorithms used for the algorithm.

### Naïve Bayes

Naïve Bayes classified data based on the highest probability of each class. The presence of each feature in a class is independent of other features in the same class (MathWorks, 2024a).

### k-Nearest neighbour (kNN)

KNN will categorise objects based on the nearest neighbour in the class. This algorithm assumes that objects near one another are identical (MathWorks, 2024a).

### Support vector machines (SVM)

SVM classification will try to draw a straight line to separate data points from different groups. The most significant margin gap between each data point class will be considered the best hyperplane. If a straight line cannot be drawn in between to separate the data points from different classes, then the model will lose a penalised point (MathWorks, 2024a). This model can handle small datasets with more generalisation ability than other models (Baashar et al., 2021).

### Decision tree

The tree in the label consists of a branching condition in which the predictor's value will be compared with the trained weight. The number of branches and weightage are labelled during the training process. This model will predict the responses based on the decision from the root to the leaf node. Modifying the model can be done to reduce the prediction time (MathWorks, 2024a). This model is quickly adapted as it is more precise and more straightforward to discover the predictor (Baashar et al., 2021).

**Random forest**

Random forest is an ensemble model in which this is a combination of different decision trees. This combination aims to ensure that the model can generate better results. Therefore, the model started with a decision tree. While travelling down the branches, the data will be separated into different branches based on specific variables.

**Ensembles models**

Ensembled models mean combining a few weak models to become a better model. All the models will be trained independently with a bootstrap sample. Boosting means adding or adjusting the weightage, which shows a weaker result. This action will help emphasise the wrongly classified observation or train the model with data with a minimised mean-squared error between the observed response and the predicted outcome from all the previous training sets.

**Logistic regression**

Logistic regression is a classification algorithm that can only provide binary results like zero and one as the prediction result. This model predicts the outcome based on the probability of occurrence in the training dataset (Wakefield, 2024).

**Gradient boosting model (GBM)**

This is a boosting ensemble model. This model will combine different models to create a better prediction result. Besides, the model will also incorporate multiple weak or average predictors as a booster to the model (S. Ray, 2024).

**Neural networks**

The neural means neurons, and networks mean linkage. Therefore, neural networks are neurons that link like a network. Each neuron is in a layered structure that connects the inputs to the desired output. The model will be trained many times to allow the strengths of linkage between neurons to be modified. This ensures the prediction result has the highest accuracy (MathWorks, 2024a). This model can help with data with nonlinear and complex relationships between input and output data (Baashar et al., 2021).

**Multilayer perceptron (MLP)**

Multilayer Perceptron is a multi-layer neural network. The neurons between the input and out layers are also known as the hidden layer. The model will be called a deep neural network when there is more than one hidden layer. During model preparation, the hidden layer is tuned using cross-validation techniques. The cross-validation technique is used to find the correct weightage for each of the predictors in the model (Banoula, 2023).

**Deep learning**

This is the model using neural network architecture. This is also called deep neural networks. Deep learning models work very well with large datasets in which the dataset comes with labelled data. A deep neural network can support up to 150 hidden layers. Deep learning has the speciality of learning features from an extensive dataset without needing to extract them manually. This model can support ongoing improvement when the size of the dataset increases (MathWorks, 2024b). Thirty-one research studies were found preparing a model to get the highest accuracy in predicting students' academic performance. Eighteen research studies have tried to find the best model among all the

commonly used models. At the same time, five other studies focus solely on one model. Still, there are additional processes before model preparation to improve the data quality and increase accuracy. Table 9is the list of the paper, which is a comparison between a commonly used algorithm and the result of finding the model with the highest accuracy in predicting students' academic performance.

**Table 9: List of the highest accuracy among the models in the paper**

| **Highest Accuracy Model** | **Reference** |
|---|---|
| Random Forest | Owusu-Boadu et al., 2021; Yağcı, 2022; Zhang et al., 2021 |
| Generalized Linear Module (GLM), Gradient Boosted Tree (GBT) | Inusah et al., 2022 |
| Linear Regression | Yohannes & Ahmed, 2018 |
| Neural Network | Baashar et al., 2021; Feng, 2019; Veluri et al., 2022; Yağcı, 2022 |
| Naïve Bayes | Morilla et al., 2020 |
| Logistic Regression | Hashim et al., 2020 |
| Decision Tree | Huynh-Cam et al., 2022; Issah et al., 2023; Menon & Janardhan, 2020 |
| Support Vector Machine (SVM) | Bayirli et al., 2023; Huynh-Cam et al., 2022 |
| Multilayer Perceptron | Huynh-Cam et al., 2022; Jalota & Agrawal, 2019 |
| Weighted voting classifier | Zulfiker et al., 2020 |

In the comparison done by Owusu-Boadu et al. (2021) among four different algorithms, the random forest is the model that provides the highest accuracy (0.765) in the result. According to Zhang et al. (2021), who compared six different models for four subjects in China's institution and public dataset, random forest outperformed another model even though it is another subject and in a separate dataset. These align with Yagci (2022), which reported the highest accuracy (0.75) among the six algorithms. Inusah et al. (2022) compare the Generalized Linear Module (GLM) and the Gradient Boosted Tree (GBT). Both models produced the prediction with the same accuracy of 84.5%. Neither model differs in precision, f-measure, sensitivity, or specificity value. Yohannes & Ahmed's (2018) research compares three different algorithms, and linear regression turns out to be a slightly better model, having a correlation coefficient value of 0.98, which is identical to the support vector machine but has a slightly different of 0.0005 of square root mean square deviation (RMSE) and faster in processing time. There are a total of four research studies that achieved the highest accuracy in neural networks. Feng (2019) reports that the model can produce a report with as high as 97.6% accuracy. In contrast, Veluri et al. (2022) said that artificial neural networks can give predictions with up to 94% accuracy. Although Yagci (2022) reported that random forest and neural networks have the same accuracy rate, the neural networks have a slightly better AUC, F1, precision, and recall rate. Baashar et al. (2021) are doing a systematic review by categorising previous research between 2015 and 2019 and concluded that the artificial neural network has the highest accuracy. Only Morilla et al. (2020) reported that Naïve Bayes had the highest accuracy, 73.61%. The only research that stated that Logistic Regression provides the highest accuracy among the three algorithms being applied is highlighted by Hashim et al. (2020). The research predicts a student's final grade and final status (passed or failed), showing an accuracy rate of 68.7% and 88.8% respectively. Four research studies mentioned that the Decision Tree gives the highest accuracy among all the algorithms included in the research. Zulfikar et al. (2020) and Menon & Janardhan (2020) reported that the accuracy rate for the decision trees is 0.81 and 0.96, respectively. Huynh-Cam et al. (2022) predict results for regular and disabled students. On accuracy for ordinary students, there are three different algorithms: Decision Tree, Support Vector Machine, and Multilayer Perceptron, all reporting 100% accuracy. Issah et al. (2023) were doing a systematic review of the last seven years and concluded that fourteen researchers highlighted that the decision tree produced the highest accuracy result. Bayirli et al. (2023) stressed that the Support Vector Machine has an accuracy rate of 73%, the highest among the three algorithms discussed in the research. Jalota & Agrawal (2019) reported that 76.07% is the highest accuracy rate obtained from Multilayer Perceptron among the five algorithms. Zulfiker et al. (2020) found that the Weighted Voting Classifier produces the highest accuracy rate of 81.73%

among the seven algorithms. Table 10 is a list of literature reviews which focus only on one type of algorithm while preparing the predictive model for student academic performance.

**Table 10: List of a literature review focusing on one algorithm in predicting student's academic performance.**

| Reference | Model |
|---|---|
| (Hamoud et al., 2018) | Random Tree |
| (S. Li & Liu, 2021) | Deep Neural Network |
| (Yakubu & Abubakar, 2022) | Logistic Regression |
| (Dabhade et al., 2021) | Support Vector Regression |

Likewise, the eighteen research papers mentioned above and the five other studies focus on only one algorithm. Hamoud et al. (2018) aim to explore and test the entire process using random trees from variable selection, applying algorithm and result evaluation. The whole process uses an algorithm under the same random tree algorithm. Li & Liu (2021) focus on the Neural Network Algorithm end-to-end process. In between the process, the researcher adds in extra processing like data initialisation and preprocessing to increase the accuracy of the result. In line with Li & Liu (2021), Lau et al. (2019) used the same algorithm to evaluate predictors on different groups of students. On the other hand, Yakubu & Abubakar (2022) consider different sets of predictors but with logistic regression while building the model. Likewise, Dabhade et al. (2021) use Support Vector Regression to evaluate the predictor, and before building up the model, detailed analysis is carried out to improve the academic performance prediction result.

## 7. Hyperparameter tunning and model testing

Hyperparameter tuning is fine-tuning the external configuration variables of the model. Different variables are available based on different models that need to be tuned. Hyperparameter tuning improves the model's accuracy to ensure the model gets the optimal result. There are five commonly used hyperparameter tunning techniques (AWS, 2024).

### Grid search

This approach lists all possible combinations of hyperparameters in a grid format. Each combination is being tested out and recorded. This approach is time-consuming and consumes much processing power as a long list of combinations can be tested. However, this will be the best to find the best combination of hyperparameters (GeeksforGeeks, 2023).

### Random search

A random search randomly selects a predetermined set of combinations of hyperparameters. It is repeatedly using the configuration to execute the data with the model and log the performance of the model. This approach will reduce unnecessary computation and help to overcome the drawbacks mentioned in grid search (GeeksforGeeks, 2023).

### Bayesian optimisation

Bayesian optimisation considers the previous evaluation result before starting a new iteration. This approach will help to cut down the time spent on unsuitable hyperparameters. The next execution iteration is always a better combination than the previous iteration (Agrawal, 2020; GeeksforGeeks, 2023).

### Successive halving search

This is an incremental testing. All sets of hyperparameters are given an equal budget time, and a limited amount of data is tested against all sets of hyperparameters. Upon completion of training of all models, half of the worst performance hyperparameters will be removed from the list. In the next training iteration, more data will be added to train all the models and remove half of the worst performance hyperparameters. The iteration will be repeated until the best set of hyperparameters is found (Agrawal, 2020).

### Hyperband search

Hyperband search is an extended version of successive halving search. The fixed budget remains unchanged in hyperband search and changes the hyperparameter value. When the value of the hyperparameter increases, it will cause the budget time to reduce; this will make the process stop executing. While the outer layer of the loop changes the hyperparameter value, the inner layer performs successive halving searches to get the best sets of hyperparameters (Agrawal, 2020). Table 11 shows the list of previous studies that use a different type of hyperparameter tuning.

**Table 11: List of literature review for hyperparameter tunning**

| Reference | Model | Hyperparameter Tunning |
|---|---|---|
| (Prochukhan, 2023) | Neural Network | Hyperband Optimization |
| (Zhou et al., 2021) | Neural Network | Bayesian Optimization |
| (Ghate & Hemalatha C, 2023) | Support Vector Machine (SVM) | Grid Search |
| (L. Wu et al., 2022; Yi & Bui, 2021) | Deep Learning | Bayesian Optimization |
| (Kaur et al., 2020) | Deep Learning | Grid Search |
| (Kaensar & Wongnin, 2023) | Random Forest | Random Search |

There is hardly found any research on predicting student academic performance using hyperparameter tuning. This might be due to the reason that the academic field does not fall under the critical system domain that requires high accuracy in predictive modelling. There are hyperparameter tuning in other research fields like medicine, ocean engineering and science, cybersecurity and traffic prediction. In between, medical is 50% of all research implementing hyperparameter tuning. The deep learning model is the most complicated, standing over half of the previous research and implementing hyperparameter tunning. Besides, most deep learning models use Bayesian Optimization hyperparameter tunning; only one research uses grid search. The researcher using grid search also highly recommends using Bayesian Optimization as resources will not be wasted on the poor performance hyperparameter. Prochukahn (2023) also highlighted that hyperband optimisation is more efficient in terms of time-consuming compared to other approaches.

Kaensar & Wongnin (2023) is found to be the only research paper that performs hyperparameter tunning to get the best model to predict the performance of first-year students and student performance for each program in the faculty. The researchers randomly select hyperparameters to try out with the model. The selection of using random search is supported by Esmaeili et al. (2023) which focusing comparing different types of hyperparameter tuning. The research highlighted that the random search method can be implemented in machine learning hyperparameter tunning (Esmaeili et al., 2023).

## 8. RESEARCH METHODOLOGY

Figure **7** is illustrating the way research methodology that the research is carried out.

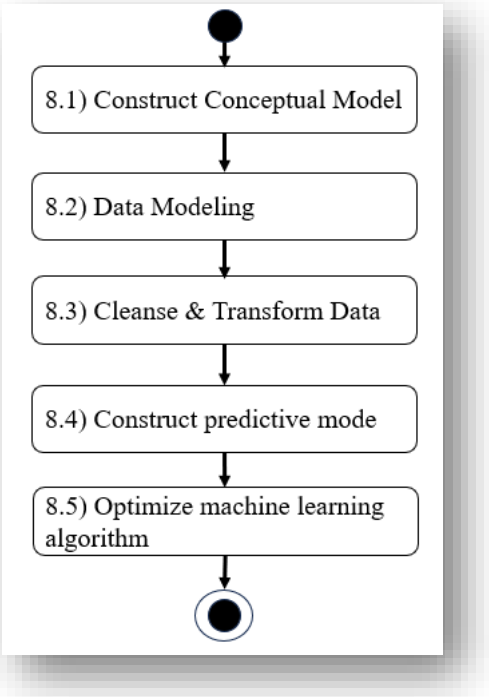


**Figure 7: Research methodology**

### 8.1 Construct conceptual model

This research aims to identify factors that affect students' academic performance. Factors and student's academic performance are considered variables as it varies from time to time within this research. The Independent variable is the value that will influence the other variable. On the other hand, the dependent value is the value that independent variables will influence. The predictors are the independent variable, whereas the student's academic performance is the dependent variable.

The Program for International Student Assessment (PISA) is research from the Organisation for Economic Co-operation and Development (OECD) to look for the need for cross-national comparable evidence on student performance. The PISA dataset is selected as it covers the information related to the student's backgrounds, attitudes toward learning, and some of the main factors affecting their home and school learning. With the help of the PISA dataset, more accurate factors can be identified from the four aspects involved during the education process: students, family, teacher, and school (PISA, 2020b).**Error! Reference source not found.**Figure 8 illustrated the factors (independent variables) have been categorised into four main categories: student-related, family/parents-related, teacher-related, and school-related. Based on PISA 2018 datasets, there are a total of 925 questions (student dataset) that have been filled up by students and parents, 180 questions (school dataset) that have been given to school principals related to school, and 333 questions (teacher dataset) has been shown to teachers (separated into teachers who teach language and another teacher). The complete list of questionnaires refers to the PISA website (PISA, 2020a).

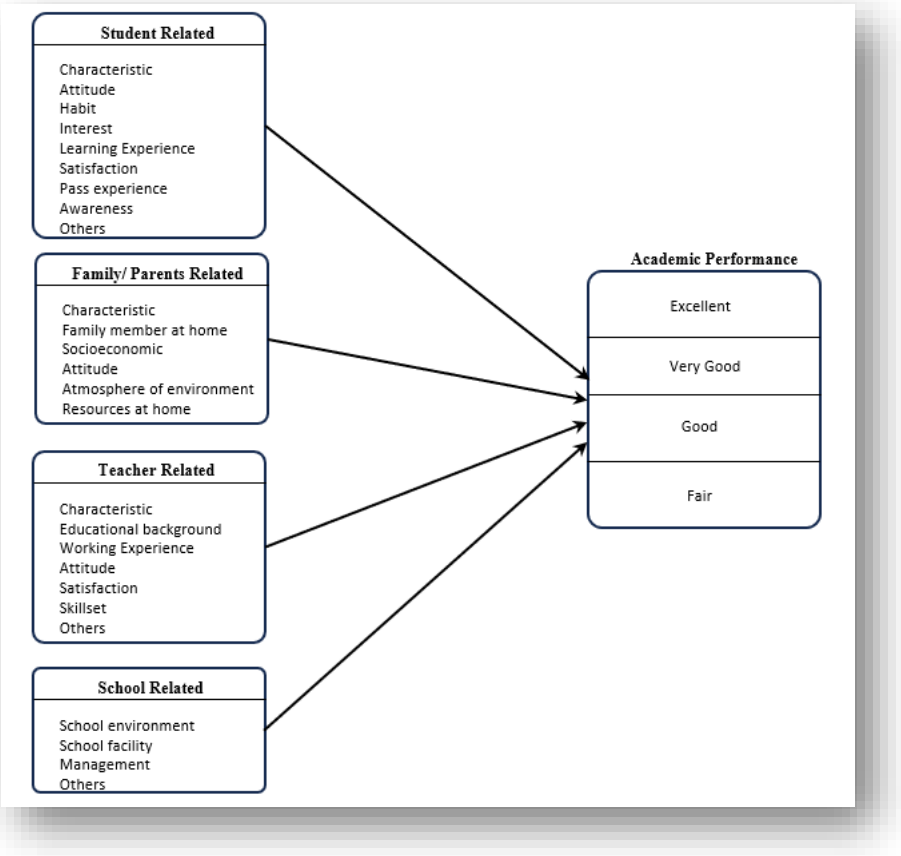


**Figure 8: Independent variables and dependent variables**

The variables are then grouped into four categories: student-related, family/parents-related, teacher-related, and *school-related* variables. These variables are categorised to identify the main factor that carries the highest weightage and impacts academic performance from which aspect.

Following is the sample of the variable categorised into each aspect accordingly:

Table 12 shows some samples of *student-related* variables being grouped into different sub-categories.

**Table 12: Student-related variable**

| **Subcategory** | **Questionnaire** |
|---|---|
| *Characteristic* | - Adaptation of instruction<br>- Age<br>- Country of Birth National Categories |
| *Attitude* | - Student's attitudes towards immigrants<br>- Attitude towards school: learning activities<br>- Perceived autonomy related to ICT use |
| *Habit* | - Joy/Like reading<br>- Learning time (minutes per week) |

Table 13 shows some samples of *family/parents-related* variables being grouped into different sub-categories.

**Table 13: Family/parents-related variable**

| Subcategory | Questionnaire |
|---|---|
| *Socioeconomic* | - Index of father's occupation<br>- Index of mother's occupation<br>- Father's educational level |
| *Attitude* | - Parent's attitudes towards immigrants<br>- Parental support for learning at home<br>- Parent's emotional support |
| *Resources at home* | - ICT use outside of school<br>- ICT available at home<br>- Family wealth |

Table 14 shows some samples of *teacher-related* variables being grouped into different sub-categories.

**Table 14: Teacher-related variable**

| Subcategory | Questionnaire |
|---|---|
| *Educational Background* | - Subject included in initial training<br>- Originally trained teacher |
| *Attitude* | - Teacher-directed instruction<br>- Teacher's self-efficacy in maintaining positive relations with students |
| *Skillset* | - Perception of cooperation at school<br>- Teacher support in the language lessons<br>- Teacher's use of specific ICT applications |

Table 15 shows some samples of *school-related* variables being grouped into different sub-categories.

**Table 15: School-related variable**

| Subcategory | Questionnaire |
|---|---|
| *Management* | - Class size<br>- Creative extra-curricular activities<br>- Shortage of educational staff |
| *School facility* | - Number of available computers per student<br>- Propotion of available computers that are connected to the internet |

**GRADE** (academic performance) is the independent variable in this study. GRADE is a relative index computed using variation between countries. GRADE is a derived value from ST001 from the student questionnaire defined according to the country (PISA, 2020c). The GRADE is converted into four categories: Excellent (3,2), Very Good (1,0), Good (-1, -2), and Fair (-3, -4).

### 8.2 Data modeling

The PISA 2018 dataset contains records from seventy-nine (79) countries, including six hundred thousand (600,000) students, one hundred seven thousand three hundred sixty-seven (107,367) teachers and twenty-one thousand nine hundred and three principals of school (21,903) who responded and participated in the questionnaire and assessment. The information is stored separately in three datasets: student, teacher, and school. To link all the datasets into one big dataset, a star schema is implemented to put all the needed variables into a fact table with the help of IBM SPSS software. After the process of combination, analysis of data is carried out. Those records which have more than 50% empty fields are being removed from the dataset. There are three hundred and

twenty-two thousand eight hundred three nine (322,839) records being removed. There are almost two million (1,928,469) records in the dataset.

During the process of variable selection, there are one hundred and eight (108) variables being removed from the dataset. Eight variables are reference data, ninety-one are dependence variables, eight variables are derived from another derived column and only one variable has more than 50% missing value.

Table 16 show the summary of the total variable being removed during the variable selection process.

**Table 16: Summary of removal variable**

| **Reason of Removal** | **No. Variable** |
|---|---|
| Reference data | 8 |
| Dependence variable | 91 |
| Derived from another derived column | 8 |
| More than 50% missing value | 1 |
| **Total:** | 108 |

Before the variable selection takes place, the dataset contains one hundred and twelve (112) derived columns from the student dataset, forty-nine (49) from the teacher dataset and nineteen (19) from the school dataset. After the selection process, there are seventy-four (74) derived from the student questionnaire, seventeen (17) from the school questionnaire, and forty-eight (48) from the teacher questionnaire.

The Table 17 shows the original record from the three datasets after the cleansing and variable selection process.

**Table 17: Variable to include in the predictive model**

| **Dataset** | **No. Column in Dataset** | **Derived Column** | **After Selection** |
|---|---|---|---|
| **Student** | 208 | 112 | 80 |
| **Teacher** | 53 | 49 | 48 |
| **School** | 23 | 19 | 16 |

### 8.2 Cleanse and transform data

### Independent variable data transformation

All collected answers from the questionnaires are transformed into the derived column. The following questions are related to students which are derived into the variable ATTLNACT (Attitude towards school: learning activities):

ST036Q05TA - Thinking about your school: Trying hard at school will help me get a good job.

ST036Q06TA - Thinking about your school: Trying hard at school will help me get into a good college.

ST036Q08TA - Thinking about your school: Trying hard at school is essential.

The following questions are related to family/parents and are derived into variable JOYREADP (Parents' enjoyment of reading):

PA158Q01HA - Statements about reading? I read only if I have to

PA158Q02IA - Statements about reading? Reading is one of my favourite hobbies

PA158Q03HA - Statements about reading? I like talking about books with other people

PA158Q04IA - Statements about reading? For me, reading is a waste of time

PA158Q05HA - Statements about reading? I read only to get the information that I need

The following questions are teacher-related and are derived into variable GCSELF (Teacher's self-efficacy in multicultural environments).

TC209Q01HA - Teach in multicultural environments: I can cope with the challenges of a multicultural classroom.

TC209Q02HA - Teach in multicultural environments: I can adapt my teaching to the cultural diversity of students.

TC209Q05HA - Teach in multicultural environments: I can ensure that students with and without migrant backgrounds work together.

TC209Q06HA - Teach in multicultural environments: I can raise awareness for cultural differences amongst the students.

TC209Q09HA - Teach in multicultural environments: I can contribute to reducing ethnic stereotypes between the students.

The following questions are school-related and are derived into variable STAFFSHORT (Shortage of educational staff).

SC017Q01NA - School's instruction hindered by: A lack of teaching staff.

SC017Q02NA - School's instruction hindered by: Inadequate or poorly qualified teaching staff.

SC017Q03NA - School's instruction hindered by: A lack of assisting staff.

SC017Q04NA - School's instruction hindered by: Inadequate or poorly qualified assisting staff.

Independent variable data cleansing

One hundred forty (140) columns contain missing data in the column which is less than 50%. There are one hundred thirty-eight (138) variables being replaced with an average value and two (2) variables replaced with zero (0). There is no field which is available to replace with majority value in the dataset.

Table 18 shows a summary of the dataset being cleansed.

**Table 18: Summary of the dataset being cleansed**

| **Way of cleansing** | **No. Variable** |
|---|---|
| Replace with the average value | 138 |
| Replace with majority value in the dataset | 0 |
| Replace value with zero | 2 |

## Imbalance dataset

Figure 9 shows the distribution of the number of students into the grade categories.

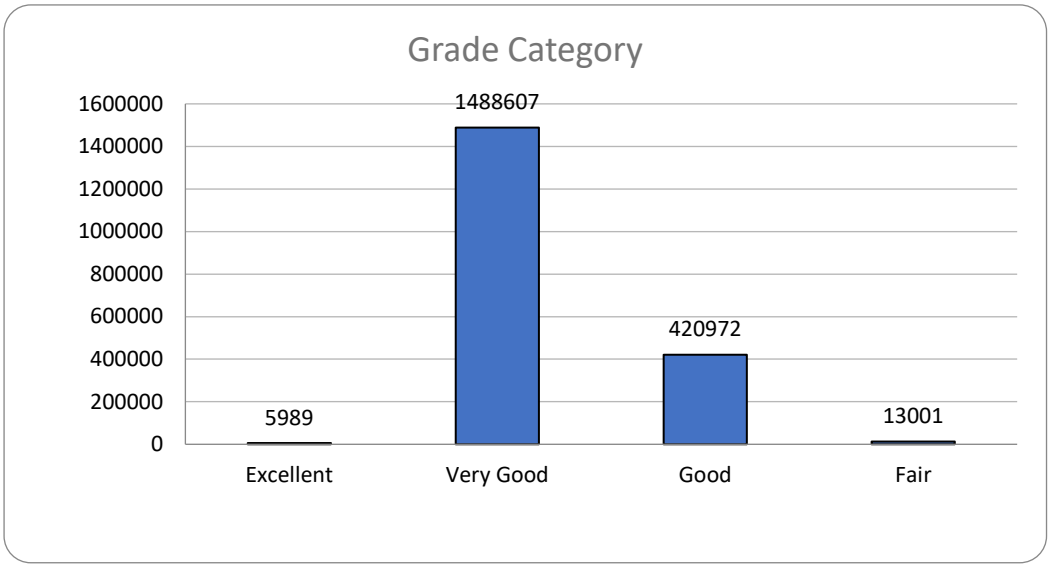


**Figure 9: Imbalance distribution of students split into grade category**

From the chart, an imbalance classification among the grade categories can be noticed. This will cause bias to happen in the training dataset. The Synthetic Minority Over-sampling Technique (SMOTE) is implemented. The "Excellent" and "Fair" are over-sampling to fill the gap with "Good" and down-sampling of "Very Good" to be the same as "Good".

Figure 10 shows the data distribution after SMOTE was implemented into the dataset.

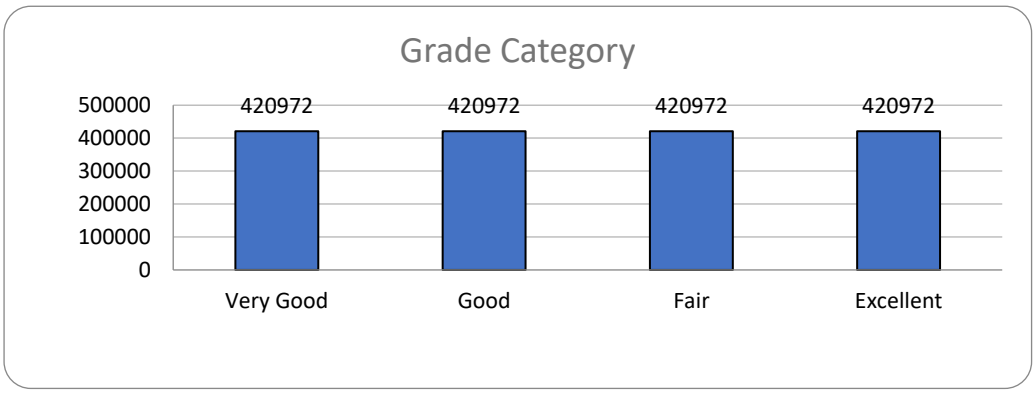


**Figure 10: Distribution of students after SMOTE**

After pre-processing the dataset, the dataset is ready to feed into the model to test for accuracy and classification error. With the help of RapidMiner, Table 19 shows the accuracy and classification error results for eight standard machine learning models.

**Table 19: Result of accuracy and classification for all the model**

| Model | Accuracy | Classification Error |
|---|---|---|

| | | |
|---|---|---|
| **Generalised Linear Model** | 77.3 | 22.7 |
| **Random Forest** | 60.8 | 39.2 |
| **Decision Tree** | 43 | 57 |
| **Fast Large Margin** | 25 | 75 |
| **Logistic Regression** | 24.9 | 75.1 |
| **Naïve Bayes** | 52.2 | 47.8 |
| **SVM** | 25.1 | 74.9 |
| **Deep Learning** | **90.1** | **9.9** |

From Table 19, Deep learning has the highest accuracy (90.1) and lowest classification error (9.9). Deep Learning is the chosen model to continue for hyperparameter optimisation. Random Search is the hyperparameter optimisation method implemented for this research. Python is the programming language used for the process of hyperparameter optimisation. The library that is imported during the model preparation process is Pandas. The changed hyperparameters are data split, activation function, random seed and epochs. Table 20 shows the difference between hyperparameters used in the auto model from Rapidminer and Panda in Python.

**Table 20: Different hyperparameters used in Rapidminer and python**

| Hyperparameter | Rapidminer | Python |
|---|---|---|
| **Data Split** | 0.9 | 0.7 |
| **Activation Function** | Rectified Linear Unit (ReLU) | Hidden Layer - ReLU<br>Output Layer - Softmax |
| **Random Seed** | 1992 | 42 |
| **Epochs** | 10 | 50 |

Data split is the portion of the data split into training and test data. The test data remains real-world data to allow for the validation of the predictive data. Improper ways of splitting data will create bias in the model and cause overfitting of the model (Yadanar Lin, 2024). The data split in Rapidminer is 0.9, in which 90% is the training data and 10% is the test data. On the other hand, Python is 0.7, in which 70% is the training data and 30% is the test data.

The activation function adds weightage to the input of the data to allow the data to have the characteristic of a non-linear function. This allows the activation of an artificial neuron (Oppermann, 2022). In Rapidminer, ReLU is the activation function being implemented. ReLU is the action function that rectifies the value when the value is greater than zero. Figure 11 shows the mathematical definition of the ReLU function.

$$R(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

**Figure 11: Mathematical definition of ReLU function.**

On the other hand, Pandas implemented ReLU for the hidden layer of the deep learning model and Softmax for the outer layer. Softmax is the activation function that will make the input values fall between zero (0) and one (1). The total of all input will be a sum of one (1). Figure 12 is the mathematical definition of the Softmax function.

$$y_j = softmax(z_j) = \frac{e^{z_j}}{\sum_{j'} e^{z_{j'}}}$$

yj: The probability of value J

zj: The given value from Softmax

e: The base of the natural logarithm (approximate value 2.71828)

$e^{zj}$ : Computation of exponential of each input value for zj

$\sum_{j'} e^{zj'}$: Summation of all the exponentials value from all inputs for j′ iteration.

**Figure 12: Mathematical definition of Softmax**

The random seed ensures the model is divided the same way every time the data training is done. The model uses as much data as possible (Bansal, 2020). Rapidminer is using 1992, whereas Panda is using 42. The random value is consistently used across every training iteration (Lee, 2024).

Epochs are the number of iterations of training and testing of the model (Simplilearn, 2023b). In Rapidminer, ten (10) rounds for Pandas are fifty (50) rounds.

After hyperparameter tuning, the accuracy of the model reached 100%. Figure 13 is the snippet of the result.

```
"Epoch 50/50\n",
"29468/29468 [==============================] - 38s 1ms/step - loss: 5.4760e-04 - accuracy: 0.9999 - val_loss: 2.7325e-04 - val_accuracy: 1.0000\n",
"15787/15787 [==============================] - 12s 732us/step\n",
"              precision    recall  f1-score   support\n",
"\n",
"           0       1.00      1.00      1.00    126461\n",
"           1       1.00      1.00      1.00    125823\n",
"           2       1.00      1.00      1.00    126735\n",
"           3       1.00      1.00      1.00    126148\n",
"\n",
"    accuracy                           1.00    505167\n",
"   macro avg       1.00      1.00      1.00    505167\n",
"weighted avg       1.00      1.00      1.00    505167\n",
"\n"
```

**Figure 13: Result after hyperparameter tuning in Python**

## 9. CONCLUSION

### 9.1 To cleanse and transform three significant data sources from PISA datasets (student, school and teacher) into one dataset.

Three data sources are 600,000 records in student datasets, 100,000 in teacher datasets and 21,000 in school datasets. Training datasets should not be left blank. Therefore, each dataset undergoes a cleansing process by replacing it with average, majority, and zero values. The independent variables are transformed using simple indices and item response theory (IRT). On the other hand, the dependent variables are categorised into four categories: Excellent, Very Good, Good and Fair. After cleansing and transformation, variable selection is carried out. The variables referencing data, the dependence variable, the derived value from another derived column, and the variable with more than 50% missing values removed are not considered part of the variables. The original two hundred eighty-four (284) variables are reduced to one hundred seventy-six (176) variables, which is one-hundred and three (103) variables from student datasets, fifty-two (52) variables from teacher datasets and twenty-one (21) variables from school datasets. After cleansing, transforming, and variable selection, the three datasets are combined using star schema, the fact table of which is the one dataset to train the model. This created datasets containing almost two million records distributed unequally between four GRADE groups. The SMOTE technique was implemented to ensure bias did not occur in the training data. Therefore, the first objective of this research is achieved.

### 9.2 To measure the performance of popular machine learning-based predictive models.

The majority of the previous research uses SMOTE to upsize the minority class. This is to make sure the training dataset has a non-bais dataset to enable the model to produce a fair prediction. When SMOTE is used, the original pattern of the data remains. At the same time, the technique can fill the gap between the majority and minority classes. This is a highly recommended technique to be applied when imbalance classification happens.

### 9.3 To measure the performance of popular machine learning-based predictive models.

Out of eight machine learning models, deep learning outperformance has 90.1% accuracy and 9.9 classification errors. This follows with a generalised linear model (accuracy: 77.3%, classification errors: 22.7) and random forest (accuracy: 60.8%, classification errors:39.2). Therefore, the second objective of this research is achieved.

### 9.4 To construct a machine learning-based predictive model.

Deep learning is selected as the baseline of the predictive model. With a random use of data split 0.7 and activation function for the hidden layer using ReLU and output using softmax and random seed set at 42 and epochs 50, the performance of this model is further improved to hit 100% accuracy.

Cleansed and transformed data has been successfully combined into a valuable dataset to train all the classification models to identify the highest accuracy model. After comparing the models, deep learning was successfully recognised as the best predictive model and is being improved to increase accuracy. There is another point of view that the preparation can be done. The current SMOTE is implemented for both training and testing data. However, based on the suggestion from Ozbun (2021), SMOTE should only implemented for training data, not testing data. This is to prevent the real-world data from being tempted and remain the origin of the testing data so that the predicted result can be verified with actual real-world data (Ozbun, 2021). A comparison can be made to check the accuracy of the predictive model using different ways of implementing SMOTE on imbalanced datasets. Explainable AI (XAI) can be used to understand the prepared predictive model and each part of the activity so that each part of the model can be further analysed and improvised by removing possible bias (IBM, 2024). Therefore, the third objective for this research is achieved.

In conclusion, all the objectives were achieved. However, whatever has been done so far is to determine the student's possible academic performance. After knowing the students' academic performance, action should be taken to help the student before the following assessment. The reports on the root cause of poor academic performance have to be produced so that the model can propose the possible action to be taken by the student, family members, teachers or schools so that the student's academic performance will be better than predicted. This can be part of the custom-made service based on the student's situation. The educational organisation can improve the quality of service by ensuring the quality of the graduates. This will indirectly help the country to have a quality workforce, which can push the country's development to a higher level and ensure the quality of life within the country can be improved.

## REFERENCES

Abdullah, N. A., Shamsi, N. A., Jenatabadi, H. S., Ng, B. K., & Mentri, K. A. C. (2022). Factors Affecting Undergraduates' Academic Performance during COVID-19: Fear, Stress and Teacher-Parents' Support. *Sustainability (Switzerland)*, *14*(13). https://doi.org/10.3390/su14137694

Agnafors, S., Barmark, M., & Sydsjö, G. (2021). Mental health and academic performance: a study on selection and causation effects from childhood to early adulthood. *Social Psychiatry and Psychiatric Epidemiology*, *56*(5), 857–866. https://doi.org/10.1007/s00127-020-01934-5

Agrawal, T. (2020). Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient. In *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-6579-6

Akiode, A. (2020, July 10). *How To Carry Out K-Fold Cross-Validation On An Imbalanced Classification Problem*. Analytics Vidhya. https://medium.com/analytics-vidhya/how-to-carry-out-k-fold-cross-validation-on-an-imbalanced-classification-problem-6d3d942a8016

Aman, F., Azhar Rauf, Rahman Ali, Farkhund Iqbal, & Asad Masood Khattak. (2019). *A Predictive Model for Predicting Students Academic Performance*.

American Psychological Association (APA). (2022, September). *Socioeconomic status*. American Psychological Association (APA). https://www.apa.org/topics/socioeconomic-status

Amin, M. M., Sutrisman, A., & Dwitayanti, Y. (2021). Development of Star-Schema Model for Lecturer Performance in Research Activities. *International Journal of Advanced Computer Science and Applications*, *12*(9).

Anna Sudderth. (2022, April 7). *How Important is Education for Economic Growth?* XQ Rethink Together. https://xqsuperschool.org/rethinktogether/how-important-is-education-for-economic-growth/

Arashpour, M., Golafshani, E. M., Parthiban, R., Lamborn, J., Kashani, A., Li, H., & Farzanehfar, P. (2023). Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Computer Applications in Engineering Education*, *31*(1), 83–99. https://doi.org/10.1002/cae.22572

ARM. (2023). *What are Machine Learning Algorithms for AI?* Arm Limited. https://www.arm.com/glossary/machine-learning-algorithms

Arora, N., & Singh, N. (2017). Factors Affecting The Academic Performance of College Students. *I-Manager's Journal of Educational Technology*, *14*, 1–7.

Ashokram. (2021, January 16). *Top 10 Reasons Why Is Education Important*. https://adilblogger.com/reasons-why-education-important-society/

AWS. (2024). *What is Hyperparameter Tuning?* Amazon Web Services. https://repost.aws/knowledge-center?nc2=h_m_ma

Baashar, Y., Alkawsi, G., Ali, N., Alhussian, H., & Bahbouh, H. T. (2021). Predicting student's performance using machine learning methods: A systematic literature review. *Proceedings - International Conference on Computer and Information Sciences: Sustaining Tomorrow with Digital Innovation, ICCOINS 2021*, 357–362. https://doi.org/10.1109/ICCOINS49721.2021.9497185

Ballotpedia. (2023). *Academic performance*. https://ballotpedia.org/Academic_performance

Banoula, M. (2023, August 10). *An Overview on Multilayer Percepton (MLP)*. Simplilearn. https://www.simplilearn.com/tutorials/deep-learning-tutorial/multilayer-perceptron

Bayirli, E. G., Kaygun, A., & Öz, E. (2023). An Analysis of PISA 2018 Mathematics Assessment for Asia-Pacific Countries Using Educational Data Mining. *Mathematics*, *11*(6), 1318. https://doi.org/10.3390/math11061318

Briones, S. K. F., Joy Dagamac, R. R., David, J. D., & Ann Landerio, C. B. (2021). Factors Affecting the Students' Scholastic Performance: A Survey Study. *Indonesian Journal of Educational Research and Technology*, *2*(1), 97–102. https://doi.org/10.17509/xxxx.xxxx

Brownlee, J. (2020). A Gentle Introduction to Imbalanced Classification. *Machine Learning Mastery: Making Developers Awesome at Machine Learning*.

Cagliero, L., Canale, L., Farinetti, L., Baralis, E., & Venuto, E. (2021). Predicting student academic performance by means of associative classification. *Applied Sciences (Switzerland)*, *11*(4), 1–22. https://doi.org/10.3390/app11041420

Cambridge University. (2023). *Cambridge Dictionary*. Cambridge University Press & Assessment.

Chakrabarty, N. (2018, October 26). *Implementation of Cluster Centroid based Majority Under-sampling Technique (CCMUT) in Python*. Towards Data Science. https://towardsdatascience.com/implementation-of-cluster-centroid-based-majority-under-sampling-technique-ccmut-in-python-f006a96ed41c

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*, 5260–5267. https://doi.org/10.1016/j.matpr.2021.05.646

Dalailama. (2023). *A Human Approach to World Peace*. The Office of His Holiness the Dalai Lama. https://www.dalailama.com/messages/world-peace/a-human-approach-to-world-peace

Dictionary.com. (2023). *Dictionary.com*. https://www.dictionary.com/browse/interests

Ebenuwa-Okoh, E. E. (2010). Influence of Age, Financial Status, and Gender on Academic Performance among Undergraduates. *Journal of Psychology*, *1*(2), 99–103. https://doi.org/10.1080/09764224.2010.11885451

Esmaeili, A., Ghorrati, Z., & Matson, E. T. (2023). Agent-Based Collaborative Random Search for Hyperparameter Tuning and Global Function Optimization †. *Systems*, *11*(5). https://doi.org/10.3390/systems11050228

Feng, J. (2019). *Predicting Students' Academic Performance with Decision Tree Predicting Students' Academic Performance with Decision Tree and Neural Network and Neural Network*. http://library.ucf.edu

García, J. D., & Skrita, A. (2019). Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees Academic Performance Based on Students´FamilyStudents´Family Environment. *Psychology*, *11*(3), 299–311. https://doi.org/10.25115/psye.v10i1.2056

GeeksforGeeks. (2023). Hyperparameter tuning. *Geeks for Geeks*. https://www.geeksforgeeks.org/hyperparameter-tuning/

Ghate, V., & Hemalatha C, S. (2023). A comprehensive comparison of machine learning approaches with hyper-parameter tuning for smartphone sensor-based human activity recognition. *Measurement: Sensors*, *30*. https://doi.org/10.1016/j.measen.2023.100925

Grant, C. (2017). *The contribution of education to economic growth*.

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26. https://doi.org/10.9781/ijimai.2018.02.004

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student Performance Prediction Model based on Supervised Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*, *928*(3). https://doi.org/10.1088/1757-899X/928/3/032019

Hoi, L. M., Ke, W., & Im, S. K. (2023). Data Augmentation for building QA Systems based on Object Models with Star Schema. *IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, 244–249.

Huynh-Cam, T. T., Chen, L. S., & Huynh, K. V. (2022). Learning Performance of International Students and Students with Disabilities: Early Prediction and Feature Selection through Educational Data Mining. *Big Data and Cognitive Computing*, *6*(3). https://doi.org/10.3390/bdcc6030094

IBM. (2023). *What is machine learning?* . IBM. https://www.ibm.com/topics/machine-learning

Injadat, M. N., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, *200*. https://doi.org/10.1016/j.knosys.2020.105992

Inusah, F., Missah, Y. M., Najim, U., & Twum, F. (2022). Data Mining and Visualisation of Basic Educational Resources for Quality Education. *International Journal of Engineering Trends and Technology*, *70*(12), 296–307. https://doi.org/10.14445/22315381/IJETT-V70I12P228

Isaac, E. (2019). *Teacher Factors Influencing Students' Academic Performance in Public Secondary Schools in Rivers State*. http://arcnjournals.org

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. In *Decision Analytics Journal* (Vol. 7). Elsevier Inc. https://doi.org/10.1016/j.dajour.2023.100204

Jalota, C., & Agrawal, R. (2019). *Analysis of Educational Data Mining using Classification*.

Kaensar, C., & Wongnin, W. (2023). Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(12). https://doi.org/10.29333/ejmste/13863

Kaur, S., Aggarwal, H., & Rani, R. (2020). Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. *Machine Vision and Applications*, *31*(5). https://doi.org/10.1007/s00138-020-01078-1

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. In *Progress in Artificial Intelligence* (Vol. 5, Issue 4, pp. 221–232). Springer Verlag. https://doi.org/10.1007/s13748-016-0094-0

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9). https://doi.org/10.1007/s42452-019-0884-7

Li, S., & Liu, T. (2021). Performance Prediction for Higher Education Students Using Deep Learning. *Complexity*, *2021*. https://doi.org/10.1155/2021/9958203

Li, Z., & Qiu, Z. (2018). How does family background affect children's educational achievement? Evidence from Contemporary China. *Journal of Chinese Sociology*, *5*(1). https://doi.org/10.1186/s40711-018-0083-8

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, *33*(3), 456–465. https://doi.org/10.7334/psicothema2021.62

Mandala, E. P. W., Permana, R., & Putri, D. E. (2021). Designing a Star Schema for Optimizing the Total Sales of Motorcycle. *Journal of Computer Scine and Information Technology*, 105–109.

MathWorks. (2024a). *What is a Machine Learning Model?* The MathWorks, Inc. https://www.mathworks.com/discovery/machine-learning-models.html

MathWorks. (2024b). *What Is Deep Learning?* MathWorks. https://www.mathworks.com/discovery/deep-learning.html

Melo, E., Silva, I., Costa, D. G., Viegas, C. M. D., & Barros, T. M. (2022). On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout. *Education Sciences*, *12*(12). https://doi.org/10.3390/educsci12120845

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

Morassi Sasso, A., Glicksberg, B. S., Böttinger, E., FreitasDa Cruz, H., Sachs, J. P., Bode, P., Datta, S., & Martensen, T. (2021). FIBER: enabling flexible retrieval of electronic health records data for clinical predictive modeling. *JAMIA Open*, *4*. https://doi.org/10.1093/jamiaopen/ooab048

Morilla, R. C., Omabe, R. D., Jane Tolibas, C. S., Edu Cornillez Jr, E. C., & Keneth Treceñe, J. D. (2020). Application of machine learning algorithms in predicting the performance of students in mathematics in the modern world. *Journal of Educational Research and Technology Management Article History*, *1*(1), 49–57.

Naka, A., Henri-Irenee Marrou, & Mehdi K. Nakosteen. (2003, September 12). *The Editors of Encyclopaedia Britannica*. Britannica. https://www.britannica.com/topic/education/additional-info#history

Natural Beach Living. (2023). *10 Reasons Why Education is Important*. An Elite CafeMedia Family & Parenting Publisher.

Nazir, A., Syafria, F., Syaputra, M. D., Gusti, S. K., & Sanjaya, S. (2022). Data Warehouse Design For Sales Transactions on CV. Sumber Tirta Anugerah. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *8*. https://doi.org/10.24014/coreit.v8i2.19800

Niu, K., Jia, B., Zhou, Y., & Lu, G. (2022). A hybrid model for predicting academic performance of engineering undergraduates. *International Journal of Modeling, Simulation, and Scientific Computing*. https://doi.org/10.1142/S1793962323500307

Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T., Misra, S., & Mbarika, V. (2021). Predicting Student Academic Performance Using Machine Learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12957 LNCS*, 481–491. https://doi.org/10.1007/978-3-030-87013-3_36

Olagbaju, O. O. (2020). Teacher-Related Factors as Predictors of Students' Achievement in English Grammar in Gambian Senior Secondary Schools. *Education Research International*, *2020*. https://doi.org/10.1155/2020/8897719

Olufemioladebinu, T., Adekunle Adediran, A., & Oyediran, W. (2018). Factors Affecting Students' Academic Performance in Colleges of Education in Southwest, Nigeria. In *British Journal of Education* (Vol. 6, Issue 10). www.eajournals.org

Owusu-Boadu, B., Nti, I. K., Nyarko-Boateng, O., Aning, J., & Boafo, V. (2021). Academic Performance Modelling with Machine Learning Based on Cognitive and Non-Cognitive Features. *Applied Computer Systems*, *26*(2), 122–131. https://doi.org/10.2478/acss-2021-0015

Ozcan, M. (2021). Factors Affecting Students' Academic Achievement according to the Teachers' Opinion. *Education Reform Journal*, *6*(1), 1–18. https://doi.org/10.22596/erj2021.06.01.1.18

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1). https://doi.org/10.1186/s13643-021-01626-4

PISA. (2020). *PISA 2018 Database*. OECD. https://www.oecd.org/pisa/data/2018database/

Polymer. (2023). *High-Dimensional Data Analysis*. Polymer. https://www.polymersearch.com/glossary/high-dimensional-data-analysis

Pusat Bahasa Abad. (2018). *Multicultural Society*. https://pbaelc.edu.my/?page_id=56

Raising Children Network (Australia). (2024). *Parents: role models and positive influences for pre-teens and teenagers*. Raising Children Network (Australia) Limited. https://raisingchildren.net.au/pre-teens/behaviour/encouraging-good-behaviour/being-a-role-model

Rajalaxmi, R. R., P. Natesan, N. Krishnamoorthy, & S. Ponni. (2019). *Regression Model for Predicting Engineering Students Academic Performance*. https://www.researchgate.net/publication/333102911

Rajendran, S., Chamundeswari, S., & Sinha, A. A. (2022). Predicting the academic performance of middle- and high-school students using machine learning algorithms. *Social Sciences & Humanities Open*, *6*(1), 100357. https://doi.org/10.1016/j.ssaho.2022.100357

Rawat, A. S. (2022, January 3). *10 Advantages of Big Data*. https://www.analyticssteps.com/blogs/advantages-big-data

Ray, M. (2023, September 18). *Sandy Hook Elementary School shooting mass shooting, Newtown, Connecticut, United States*. Britannica. https://www.britannica.com/event/Sandy-Hook-Elementary-School-shooting

Ray, S. (2024, January 24). *Top 10 Machine Learning Algorithms to Use in 2024*. Anayytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Robinson, S. K., & Robinson, K. (2022, March 2). *What Is Education For?* Edutopia. https://www.edutopia.org/article/what-education/

Rose, A. (2020). *Perceived Teacher Related Factors Influencing Academic Performance Among Public Secondary School Students In Suba Sub-country, Homa Bay Country, Kenya*.

Rossin-Slater, M. (2022, June). *Surviving a school shooting: Impacts on the mental health, education, and earnings of American youth*. Stanford University. https://siepr.stanford.edu/publications/health/surviving-school-shooting-impacts-mental-health-education-and-earnings-american

Shee, E. (2022, September 16). *Supervised vs Unsupervised Learning Explained*. SELDOM. https://www.seldon.io/supervised-vs-unsupervised-learning-explained

Simplilearn. (2023, June 6). *What is Data Modelling? Overview, Basic Concepts, and Types in Detail*. Simplilearn. https://www.simplilearn.com/what-is-data-modeling-article

Sokkhey, P., & Okazaki, T. (2020). Hybrid Machine Learning Algorithms for Predicting Academic Performance. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Issue 1). www.ijacsa.thesai.org

Suharso, W., Fardiansa, A., Munarko, Y., & Wibowo, H. (2021). Implementasi Star Schema Pada Studi Kasus Perpustakaan Berskala Universitas. *SINTECH (Science and Information Technology) Journal*, *4*(1), 1–11.

Sunil Ray. (2023, March 3). *Commonly used Machine Learning Algorithms (with Python and R Codes)*. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Tadese, M., Yeshaneh, A., & Mulu, G. B. (2022). Determinants of good academic performance among university students in Ethiopia: a cross-sectional study. *BMC Medical Education*, *22*(1). https://doi.org/10.1186/s12909-022-03461-0

Tech Target Contributor. (2024). *Star Schema*. Tech Target Network - Data Management. https://www.techtarget.com/searchdatamanagement/definition/star-schema

United Nations. (2023). *United Nations: Department of Economic and Social Affairs, Sustainable Development*. https://sdgs.un.org/goals

Veluri, R. K., Patra, I., Naved, M., Prasad, V. V., Arcinas, M. M., Beram, S. M., & Raghuvanshi, A. (2022). Learning analytics using deep learning techniques for efficiently managing educational institutes. *Materials Today: Proceedings*, *51*, 2317–2320. https://doi.org/10.1016/j.matpr.2021.11.416

Wakefield, K. (2024). *A guide to the types of machine learning algorithms and their applications*. SAS Institute Inc. .

Weaver, J. (2019). *Teacher Factors and Student Achievement as Measured by the ACT Assessment and Subsequent Teacher Perceptions of Those Factors*. https://dc.etsu.edu/etd/3540

Widyaningsih, Y., Nur Fitriani, & Devvi Sarwinda. (2019). *A Semi-Supervised Learning Approach for Predicting Student's Performance: First-Year Students Case Study*.

Wikipedia. (2022). *Wikipedia - The Free Encyclopedia*. https://en.wikipedia.org/wiki/Engineers_of_the_human_soul

Wu, L., Perin, G., & Picek, S. (2022). I Choose You: Automated Hyperparameter Tuning for Deep Learning-based Side-channel Analysis. *IEEE Transactions on Emerging Topics in Computing*. https://doi.org/10.1109/TETC.2022.3218372

Wu, Y., & Radewagen, R. (2022). 7 Techniques to Handle Imbalanced Data. *KD Nuggets*. https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes*, *51*(2), 916–934. https://doi.org/10.1108/K-12-2020-0865

Yi, H., & Bui, K. H. N. (2021). An Automated Hyperparameter Search-Based Deep Learning Model for Highway Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems*, *22*(9), 5486–5495. https://doi.org/10.1109/TITS.2020.2987614

Yohannes, E., & Ahmed, S. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications*, *180*(40), 39–47. https://doi.org/10.5120/ijca2018917057

Yusof, N. (2013). 'Not Knowing is Hence, Not Loving': The Significance of Raising Language Awareness Amongst Teachers of Content Area Learning. *GSTF International Journal on Education Vol.1 No.2*, *1*(2). https://doi.org/10.5176/2345-7163_1.2.28

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.698490

Zhou, T., Zhen, R., Hu, Q., & Hu, Z. (2021). An adaptive hyper parameter tuning model for ship fuel consumption prediction under complex maritime environments. *Journal of Ocean Engineering and Science*, *7*. https://doi.org/10.1016/j.joes.2021.08.007

Zulfiker, M. S., Kabir, N., Biswas, A. A., Chakraborty, P., & Rahman, M. M. (2020). Predicting students' performance of the private universities of Bangladesh using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, *11*(3), 672–679. https://doi.org/10.14569/ijacsa.2020.0110383