The Architecture of System for Predicting Student Performance Based on Data Science Approaches (SPPS-DSA Architecture)

Kitsadaporn Jantakun, Thiti Jantakun, and Thada Jantakoon

***Abstract*—The goals of this study are to develop the architecture of a system for predicting student performance based on data science approaches (SPPS-DSA Architecture) and evaluate the SPPS-DSA Architecture. The research process is divided into two stages: 1) context analysis and 2) development and assessment. The data is analyzed by means of standardized deviations statistically. The research findings suggested that the SPPS-DSA architecture, according to the research findings, consists of three key components: i) data source, ii) machine learning methods and attributes, and iii) data science process. The SPPS-DSA architecture is rated as the highest appropriate overall. Predicting student performance helps educators and students improve their teaching and learning processes. Predicting student performance using various analytical methods is reviewed here. Most researchers used CGPA and internal assessment as data sets. In terms of prediction methods, classification is widely used in educational data science. Researchers most commonly used neural networks and decision trees to predict student performance under classification techniques.**

***Index Terms*—Predicting student performance, data science, machine learning, SPPS-DSA architecture.**

I. Introduction

Student achievement is significant in higher education. This is because one of the criteria for a high-quality university is a solid academic record. First, there are numerous descriptions of student achievement based on previous studies [1]. Assessing learning evaluation and co-curriculum can be used to measure student progress, according to Raihana *et al*. On the other hand, most analysis shows that graduation is the most significant predictor of a pupil's success. Most universities in Malaysia have assessed the success of students utilizing their final degrees in general. Final grades are affected by course content, examination marks, final exam scores, and extracurricular activities [2]. Assessment is essential to ensure student achievement and the effectiveness of the learning process. A strategic program can be well planned by assessing a student's progress during their tenure with an organization [3]. At present, several ways of measuring student results are being proposed. One of the most widely employed methods for measuring student results is data mining. Data mining has recently gained popularity in the area of education [4]. It's referred to as educational data science. Educational data science is a technique for extracting valuable data and trends from an extensive archive of educational data [5]. Students' success can be predicted using user data and trends. As a result, educators would be better able to have an integrated instructional method.

Educators should also keep track of their students' progress. Students' learning habits could be changed, allowing the administration to enhance the system's effectiveness. As a result, data science methods can be applied to the individual needs of various organizations. Because of these problems, it is best to use a systematic analysis. The proposed systematic research aims to support the study's goals, which are:

1) to use data science approaches to create the architecture of a system for predicting student performance.
2) to assess the architecture of a system for predicting student performance.

II. Literature Review

A. AI, Machine Learning, and Data Science

The fields of machine learning, artificial intelligence, and computer science are intertwined. Unsurprisingly, they are often used interchangeably in newspapers and business correspondence, often in the same phrase. However, based on the circumstances, each of these three areas has a distinct identity. The relationship between artificial intelligence, machine learning, and data science [6]. The process of training machines to mimic human behavior, especially cognitive functions, is known as artificial intelligence [7]. Face recognition, automatic driving, and mail sorting by postal code are only a few examples. Machines have far outstripped human capacities in some ways, although we have just scratched the surface in others. Artificial intelligence encompasses a wide variety of methods, from linguistics to natural language modeling, decision science, bias, perception, robots, teaching, and so on. The desire to learn is part of the human capacity. Several other living organisms can also learn [8].

B. Data Science

Data science begins with data, which can be as plain as a few numeric observations or as complex as a matrix of millions of observations of thousands of variables [9]. Data scientists use advanced statistical techniques to uncover concrete and usable constructs within a dataset. Database

Manuscript received January 25, 2022; revised April 13, 2022.

Kitsadaporn Jantakun and Thiti Jantakun are with the Department of Computer Education, Faculty of Education, Roi Et Rajabhat University, Roi Et, Thailand (e-mail: jansri.kp@gmail.com, thiti100@gmail.com).

Thada Jantakoon is with the Department of Information and Communication Technology for Education Department, Rajabhat Maha Sarakham University, Mahasarakham, Thailand (e-mail: thada.phd@gmail.com).

doi: 10.18178/ijiet.2022.12.8.1684

systems, computational simulation, modeling, data analysis, experimentation, and business intelligence are only a few of the areas where data science and BI intersect [10]. Data science can be better understood by looking at some of its core characteristics and motives.

C. Data Science Classification

Problems in data science can be widely divided into supervised or unsupervised learning models. Supervised or guided data science attempts to infer a feature or relationship from classified training data and then apply the function to map new unlabeled data. Supervised methods predict the significance of output variables based on a set of input variables [11]. To accomplish this, a model has been developed. From a testing dataset in which the input and output values are known to be known, previously understood. The model generalizes the relationship between input and output variables and predicts a dataset of only specified input variables. The expected performance variable is often referred to as a class mark or goal variable. Supervised data science requires a large enough number of labeled records to learn the model from the data. Unsupervised or undirected data science exposes hidden patterns in unlabeled data. There are no predictability factors in unsupervised data science. This subset of data science techniques aims to uncover connections in data dependent on data point relationships [12]. Both supervised and unsupervised learners can be used in an application. Data problems in the data sciences can also be divided into such activities as grouping, regression, the study of associations, clustering, and anomalies, engine recommendations, time series prediction, and text mining [13].

D. Data Science Algorithms

An algorithm for solving a problem is a rational, step-by-step process. The plan is how a specific data issue is resolved in data science. Many learning algorithms are recursive, meaning several steps are repeated several times to achieve a restrictive state. Some algorithms have an algorithm as an input and are well-referred to as "random algorithms." A classification problem can be solved using a variety of learning algorithms, including decision trees, artificial neural networks, k-NN, and regression algorithms. The algorithm to be used is determined by the form of the data set, the purpose, the data structure, the presence of outliers, the number of records accessible, the number of attributes, and so on. In determining the output of various algorithms, the data science practitioner must first determine the algorithm (s) [14]. Hundreds of algorithms have been created in recent decades to address the challenge of data science [6].

E. Data Science Process

A data science methodology system involves the sequential progression of exploratory analysis and discovery tasks [15]. The basic data science method entails 1) identifying the challenge, 2) gathering data samples, 3) designing the model, 4) testing the model on a dataset to see how it could perform in the real world, and 5) deploying and retaining the models. Different models for the method have been proposed by numerous academic and industrial bodies over the years as data science techniques have evolved [16].

Cross-industry data mining is a leading strategy for data extraction. This data-mining model was developed using data collected collaboratively [17]. The CRISP-DM procedure is commonly used in data science technology design. The SAS Institute has established the SEMMA data processing scope for samples, explorations, changes, and evaluations. DMAIC, which is the acronym for defining, measuring, analyzing, improving, and checking, is employed in Six Sigma [18] and the fields of information administration selection, preprocessing, transformation, data mining, interpretation, and assessment [19], [20]. Since both of these systems have similar characteristics, we'll use a generalized system that strongly resembles the CRISP-DM method. A data science methodology, like every other system, suggests completing a specific series of tasks to produce the maximum results. In contrast, the retrieval of facts and expertise from documents is an iterative process. The data science method's phases are not sequential, and they often go through several cycles, back and forth between steps, often reverting to the first level to redefine the data science problem argument.

F. Decision Tree Classification in Student Performance

Kolo *et al*. [21] suggested a decision tree method for forecasting students' academic success. To improve educational quality, it is essential to be able to predict students' academic success. Students' financial status, gender, and motivation to research were discovered to influence their success. A large number of students were expected to pass, and male students had a stronger incentive to pass than female students. Hamsa *et al*. [22] developed a model for academic estimation in the electronics and connectivity and information sciences fields for students with master's degrees and bachelor's degrees using two classifying approaches for the genetic and decision tree. The predictive model that results can be used to detect student progress in any subject. Teachers can then classify students and take early measures to increase their performance over time. Since early solutions and forecasts can be made, successful results in final exams can be predicted. Raut and Nichat's [23] study uses a decision tree classification technique to assess student results. This research has largely concentrated on classification criteria used to measure effectiveness based on knowledge breadth. Provision of results and particular development needs, including student support during their learning processes and prompt decision-making to avoid academic abandonment and risk, Olaniyi *et al*. [24] propose an approach to data mining to analyze student outcomes. Data mining provides several methods for studying student performance, and task classification is used in this research to estimate student performance since there are several methods for classification involving the decision tree process. This study also looks at the accuracy of various decision tree algorithms. Hasan *et al*. [25] investigated student academic performance using a decision tree algorithm with parameters such as student behavior and academic data. The WEKA data mining method is used for the assessment of the decision tree algorithm and access period for students. The proposed analysis leads to the creation of module grades and assists stakeholders in measuring and assessing module results and

execution.

G. Neural Network Classification in Student Performance

Used artificial neural network models to predict academic success [26]. This is to rate university aspirants at different levels based on their chances of performance. This research predicts students' mistakes in different classes during their college careers and suggests strategies to avoid them. Several education scholars, according to Binh and Duy [27], have concentrated on learning styles and their recommendations. They are based on the fact that students have diverse characteristics, which lead to different behavioral patterns that influence student achievement in all subject areas. This research created an artificial neural network to forecast academic achievement based on the students' learning styles. According to Gerritsen's research [28], neural networks outperform classifiers in a broad range of data mining applications. The primary aim of this research is to decide if neural networks are a suitable classifier for predicting student performance in an LMS (learning management system) in the sense of educational data mining. The teaching characteristics are extracted from learning management system information gained during each course period and vary from usage specifics such as time spent on each page of the course to grades received for course quizzes and assignments. Okubo *et al*. [29] used a recurrent neural network algorithm to deduce students' final grades from the actual grades that were recorded in educational systems using log-based results. The log information showed students using the LMS, electronic book, and electronic portfolio systems' learning habits. This method was used in this study to log information from students and investigate prediction accuracy. Bendangnuksung and Prabu [30] suggested developing a deep neural network to forecast student performance. It is suggested that a deep neural network be used to display to students which type of category they belong to. This increases the knowledge of educational organizations, encouraging them to offer a response to vital failing students. The proposed deep neural network uses logistic regression analysis to determine whether students can pass or fail.

H. Naive Bayes Classification in Student Performance

Shaziya *et al*. [31] claim that they have established a procedure for college exams. Naive Bayes classifications have a single goal: to forecast students' final grades. A study predicting student results may be used in several different forms. By this method, teachers and students take important measures to construct naive models, and student results from the past six months are used to assess the next six months. Makhtar *et al*. [32] explore the performance of students in Sijil Pelajaran, one of the classification methods used in data mining, by using the nave Bayes classification to identify the clouded data among subjects that have influenced their performance. For the early-stage 2nd half, the nave Bayes algorithm can be used with 74% accuracy to classify students' results. The study by Patil *et al*. [33] shows that the students who choose to be engineers are progressing quickly, but the rates of abandonment are higher because of different factors and inappropriate training in India. Students aren't in a position to brightly discuss the subjects of engineering that are mathematical and complex. With the use of data mining techniques, the students' output in terms of drops and grades can be predicted. The Nauves Bayes algorithm is used in this research, and the framework can derive the most important factors affecting student performance based on the rules acquired from the established process. For the analysis, nave Bayes was applied to data that was unsupervised and focused on academic purposes by Razaque *et al*. [34] In addition to testing student knowledge, it was used as a tool for academic evaluation by both the teachers and students. When their achievements in the study became clear, students celebrated their successes. This research was conducted to determine that the students need focused intervention to minimize the possibility of failure and to implement an effective means of monitoring in the following semester. Divyabharthi and Someswari [35] developed a model for predicting student academic achievement. Since there are several classifications approaches available, this thesis employed naive classification techniques. This model can be used to make quick decisions to minimize academic risk for the student. The instructor should be aware of how well or poorly students are doing. The study looked at the relationship between test scores and socioeconomic factors to verify and improve statistical models of academic success.

I. SVM Classification in Student Performance

Asogbon *et al*. research [36], one approach for higher education organizations to achieve standard quality education is to properly predict, evaluate, and suggest instructional initiatives based on educational results. This has little impact on students who are just starting with the curriculum. A DSS focused on vector support methodology, multi-class, has been created to classify student performance in higher education institutions. The survey by Pratiyush and Manu [37] indicates that new technologies that result in huge amounts of data have arisen since the development in the education sector. Educational data mining helps facilitate the use of student resource success to forecast investment outcomes and predict new education patterns. This research examines the student placement data and classification system using a vector supporting machine for training information to find outcomes that not only help educational institutions grow their placement from acquired expertise but also improve competitiveness through data mining techniques. This study also examines the student classification method. Kadambande *et al*. [38] are developing applications for measuring student success using data mining techniques, which are now commonly used in the field of education as education has become increasingly important in the modern era. The SVM algorithm, a supervised learning system, is used in this research. The SVM algorithm is used for estimation, and the data is analyzed using regression and classification. The Support Vector Machine can assist students in determining how much further development they need to do to be eligible for placements. Oloruntoba and Akinode [39] use support vector machines to forecast students' academic progress. This study aims to investigate the connection between a student's academic profile before admission and their final academic results. The help vector

machine outperforms the majority of machine learning algorithms. The parameters of the support vector machine algorithm were modified to increase precision, and the findings reveal that the radial base function kernel with penalty outperforms the others.

J. KNN Classification in Student Performance

KNN precision is comparable with more complex techniques such as help vector machines, support trends, and kernel methods in machine learning and student performance prediction [40], [41]. Its output is comparable to, and sometimes outperforms, competing approaches, especially when extended with feature and distance weight genetic learning [42], [43]. With KNN, it is particularly well-suited to making predictions from noisy and imperfect data [44].

III. The Methodology

The research method was divided into two phases according to the research objectives. Phase 1: Investigate and categorize the tools and variables used to predict student performance. The researcher has researched the paper for synthesis, which is the analysis of knowledge about the structural synthesis of the methods and variables that are used to predict student performance. The Design of a System Architecture for Predicting Student Performance. Phase 2: A qualified evaluation of the design of the architecture for forecasting student success is a ten-qualified person assessment. The population is an expert on machine learning, data science, and student performance prediction. The samples consist of ten experts in machine learning, data science, and predicting student performance. Purposive sampling was used to select them. They have worked in these fields for at least five years and are highly experienced experts.

IV. Results

Phase 1: The study's findings in relation to research theories the architecture of a system for predicting student performance based on data science approaches competencies, the results can be seen in Fig. 1.

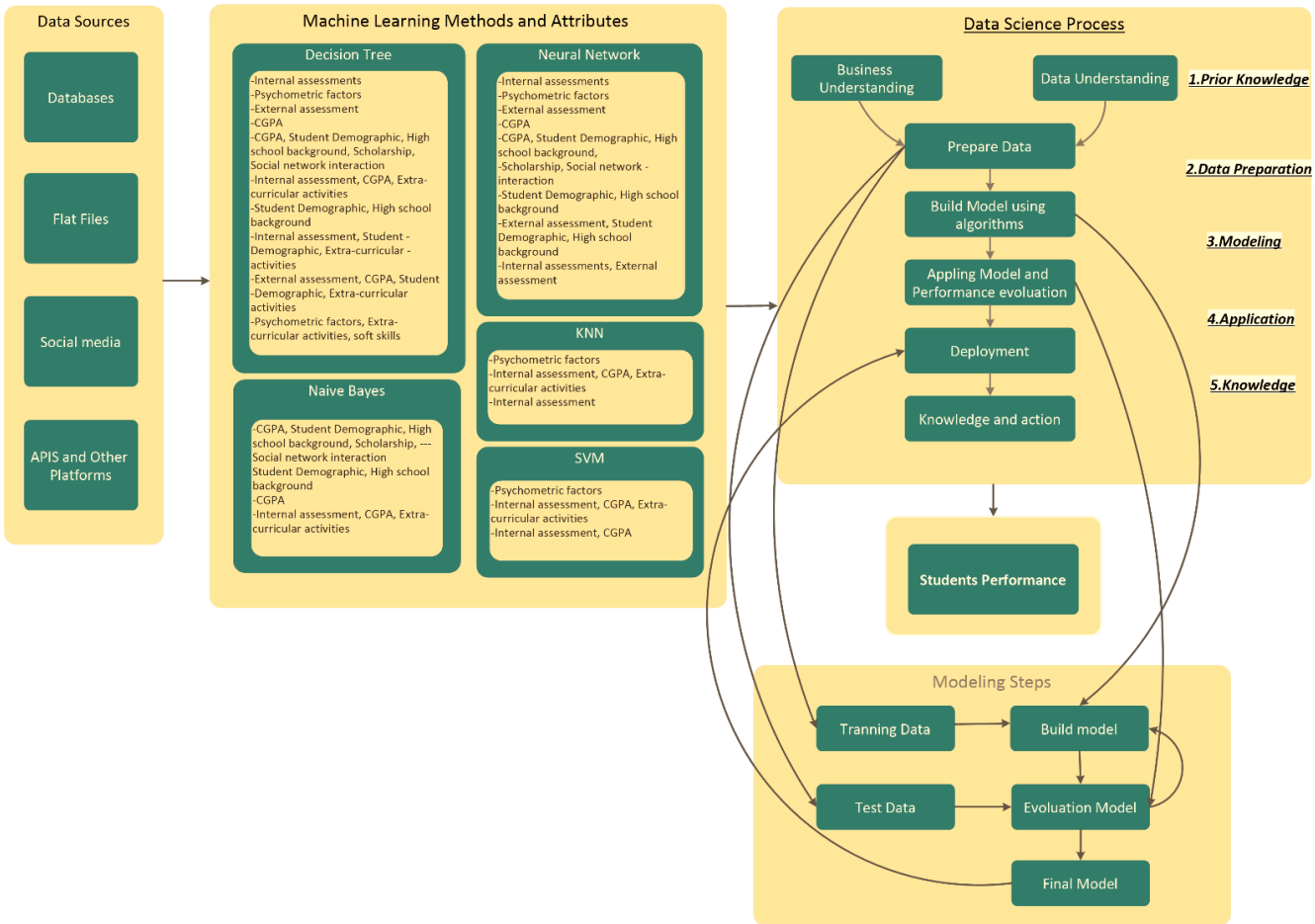

Fig. 1. The architecture of system for predicting student performance based on data science approaches (SPPS-DSA architecture).

A. Data Source

1) Databases: The program that gathers and analyzes data through communication with end-users, programs, and the database itself is known as the information management system (DBMS). The DBMS kit also contains the database management software. A database, a knowledge management platform, and the functionality that goes with them comprise an information infrastructure. The word "database" can be used to talk about either the database management system (DBMS), the database setup, or a database-related app.
2) Flat File: A flat-file database is contained inside a flat-file. There are no systems for indexing or identifying

relationships between documents, and records have a standard format. The document is straightforward. A flat file may be either plain text or binary. In this case, the data in the database can be used to make assumptions about relationships, but the way the database is set up doesn't make these relationships clear.

3) Social media has both a digital and an ideological component. It is the "collective of all internet-related applications that have a digital basis and are based on ideological and technical principles." The list of social media outlets goes on and on, but does not always include the following: social networking (Facebook or LinkedIn), photo sharing (Instagram, Photobucket, Picasa, or Picasa), video sharing (Twitch, Ustream, Google Video), and the virtual world (World of Warcraft), as these may already be a part of the description), news aggregation (StumbleUpon or Feeda).
4) APIs and other platforms Enterprises are actively involved in APIs. New architecture and operational innovations are built on these basic principles. An API supports integrations at the back-end, and APIs for front-end applications are made stable. Among them are MESH microservices management, API mediation, runtime services, data as a service, streams and event-driven APIs, pre-built back-end services, Application Connectors, and many other tools and services.

B. Machine Learning Methods and Attributes

1) A decision tree is one of the most commonly employed prediction techniques. Because of its flexibility and comprehensibility, this approach has been widely used to discover both tiny and massive data systems [45]-[47]. According to Romero *et al*., the decision tree paradigm is simple to grasp because of its reasoning process and can be explicitly converted into a system of IF-THEN laws [48]. In a web-based education framework, students' performance is assessed using features created from logged data. Among the datasets are overall grades [49], final cumulative grade point average (CGPA) [50], and marks received in specific courses [48]. Many of these databases are reviewed and evaluated to identify the most significant characteristics or variables influencing student achievement [47], [51]. The most suitable data mining algorithm would then be explored to forecast students' outcomes [52]. In their analysis [53], Mayilvaganan and Kapalnadevi contrasted classification strategies for forecasting student performance in their analysis [53]. Gray *et al*., on the other hand, looked into the predictive power of classification structures when it comes to getting into college or university [54].
2) Neural Network another common strategy in educational data mining is neural networks. One advantage of neural networks is their ability to classify all potential relations between predictor variables [54]. A neural network [55] can identify dynamic nonlinear interactions between dependent and independent variables. As a consequence, one of the most effective prediction methods is the neural network approach. As a consequence of the meta-analysis study, reports using the Neural Network method have been published. The papers [55], [56] propose a paradigm of artificial neural networks for forecasting student achievement. Admission results [57], students' attitudes toward self-regulated learning, and academic success [58] are among the principles measured by Neural Network.
3) Naive Bayes to make forecasts, researchers may also use the Naive Bayes algorithm. Make comparisons to determine the most effective prediction method for predicting student achievement [46], [52], [53], [59]. According to their results, Naive Bayes employed all of the attributes of the data. It then examined each one to demonstrate the significance and independence of each attribute [8].
4) K-Nearest Neighbor: K-Nearest Neighbor was the most accurate and performed the most. According to Bidgoli *et al*., the K-Nearest Neighbor scheme took less time to characterize students' progress as sluggish learners, good learners, solid learners, and excellent learners [53], [60]. The K-Nearest Neighbor method can accurately predict how a learner will progress through tertiary education with high accuracy [55].
5) The Support Vector Machine was chosen by Hamalainen *et al*. as their prediction tool because it works well for small datasets [61]. According to Sembiring *et al*., support vector machines have strong generalization potential and are faster than other systems [62]. Gray *et al*. found that the Support Vector Machine approach has the best prediction accuracy when it comes to determining students who are at risk of falling [54].

C. Data Science Process

1) Prior Knowledge: Prior knowledge is information on a topic that has already been learned. The data science challenge does not arise in a vacuum; it often arises on top of previously understood subject matter and contextual knowledge. In the data science phase, the previous information stage helps to figure out what the problem is, how it fits into the market, and what data is needed to solve the problem.
2) Data Preparation: The most time-consuming aspect of the project is preparing the dataset for the data science mission. Datasets that meet the requirements of data science algorithms are exceedingly unusual. Most data science algorithms enable data to be organized in a tabular format with rows of records and columns of attributes. If the data is in another format, it will need to be converted into the appropriate structure using pivot, type conversion, join, or transpose functions, among other things.
3) Modeling: A model is an explicit description of data and interactions in a dataset. Since there is insufficient quantitative evidence to use in a development scenario, a simple rule of thumb such as "mortgage interest rate decreases with rise in credit score" is a model; it provides directional knowledge by abstracting the connection between credit score and interest rate. Data science algorithms are derived from mathematics, machine learning, pattern recognition, and the body of information related to computer science, and there are a few hundred in use today. Fortunately, there are a plethora of

commercial and open-source data science tools accessible to further simplify the implementation of these learning algorithms. It is sufficient for a data science professional to have a basic understanding of the learning algorithm, how it operates, and what parameters need to be configured based on the business and data. Since they forecast an outcome variable based on one or more input variables, classification and regression tasks are predictive techniques. To learn the model, predictive algorithms need a previously established dataset. Modeling steps show the steps in the modeling phase of predictive data science.

4) Application: The stage of deployment is when the concept can be used in development. The findings of the data science process must be assimilated into the business process, which is normally done by software systems in business applications. Model readiness, technical integration, response time, model management, and assimilation are all things that need to be looked at when implementing a model.
5) Knowledge: Prior knowledge is the foundation of data science, and posterior knowledge is the incremental experience gained. The data science method, like any quantitative methodology, can reveal spurious, irrelevant trends in the dataset. Not every trend that is discovered leads to gradual awareness. It's up to the practitioner to rule out the meaningless trends and focus on the important data. In an application, the effect of data science knowledge can be calculated. It's the distinction between obtaining data through a data science method and gaining knowledge from simple data analysis. Finally, the entire data science method is a tool for asking the right questions [17] and providing feedback on how to solve a problem using the right approaches. It is meant to be used as a series of steps that help you look for information, not as a set of rules.

Phase 2: The result of appropriateness measurement and Experts' Opinion of the SPPS-DSA Architecture.

TABLE I: EXPERTS' EVALUATION OF THE SPPS-DSA ARCHITECTURE

Evaluation Lists	Level of assessment		
	$\bar{x}$	*S.D.*	*Level of suitability*
1. Data Source			
1.1 Database	4.78	0.42	Highest
1.2 Flat File	4.85	0.36	Highest
1.3 Social media	4.64	0.63	Highest
1.4 API and Other Platforms	4.71	0.46	Highest
2. Machine Learning Methods and Attributes			
2.1 Decision Tree	4.57	0.64	Highest
2.2 Neural Network	4.57	0.51	Highest
2.3 Naive Bayes	4.78	0.42	Highest
2.4 K-Nearest Neighbor	4.71	0.46	Highest
2.5 Support Vector Machine	4.64	0.49	Highest
3. Data Science Process			
3.1 Prior Knowledge	4.71	0.46	Highest
3.2 Data Preparation	4.57	0.49	Highest
3.3 Modeling	4.71	0.46	Highest
3.4 Application	4.64	0.51	Highest
3.5 Knowledge	4.64	0.49	Highest
Summary of assessment items	**4.68**	**0.50**	**Highest**

Evaluation of the overall processes of the SPPS-DSA Architecture found the mean was 4.68 and the standard deviation was 0.50. As a result, the overall processes are the highest appropriate.

V. CONCLUSION

The following is a succinct summary of the SPPS-DSA Architecture, which was established as a consequence of the expert's assessment: Among the data sources are databases, flat files, social media platforms, application programming interfaces (APIs), and other platforms. To mention a few, machine learning methods and attributes include the Decision Tree, the Neural Network, Naive Bayes, the k-Nearest Neighbor algorithm, and the Support Vector Machine. The Data Science Process comprises the following steps: prior knowledge, data preparation, modeling, application, and knowledge acquisition. The findings of the evaluation revealed that the developed architecture was the highest appropriate one for the tasks, with a mean of 4.68 and a standard deviation of 0.50. Predicting student performance is mostly useful for assisting educators and students in improving their learning and teaching processes. This paper examines prior research on predicting student performance using various analytical methods. As data sets, the majority of the researchers used cumulative grade point average (CGPA) and internal assessment. In terms of prediction techniques, the classification method is widely used in educational data science. The two most commonly used classification techniques by researchers for predicting student performance are neural networks and decision trees. Finally, the meta-analysis on predicting student performance has inspired us to conduct additional research that can be applied in our environment. It will aid the educational system in tracking students' progress in a systematic manner.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Kitsadaporn Jantakun and Thada Jantakoon developed the idea of research. Thiti Jantakun prepared the theoretical base of research and conducted a survey and statistical analysis of data.

REFERENCES

[1] T. Jantakoon and P. Wannapiroon, "System architecture of business intelligence to aun-qa framework for higher education institution," *Turkish Online Journal of Educational Technology, November Special Issue INTE*, pp. 1045-1052, 2017.
[2] Z. Raihana and A. M. F. Nabilah, "Classification of studentsbased on quality of life and academic performance by using supportvector

machine," *Journal of Academia UiTM Negeri Sembilan*, vol. 6 no. 1, pp. 45-52, 2018.

[3] J. Xu, K. H. Moon, and M. Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE J. Sel. Top. Signal Process*, vol. 11, no. 5, pp. 742–753, 2017.

[4] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art, Trans. Sys." *Man Cyber Part C*, vol. 40 no 6, pp. 601–618, 2010.

[5] D. M. D. Angeline, "Association rule generation for student performance analysis using apriori algorithm," *The SIJ Transactions on Computer Science Engineering & its Applications* (*CSEA*), vol. 1 no. 1, pp. 12–16, 2013.

[6] V. Kotu and B. Deshpande, *Data Science*: *Concepts and Practice*, Morgan Kaufmann Publishers Inc., 2019.

[7] W. Yu and F. Qun, "Application of artificial intelligence in microfluidic systems," *Chinese Journal of Analytical Chemistry*, vol. 48, issue 4, pp. 439-448, 2020.

[8] L. Yang, P. Q. Lian, Z. J. Xue, and D. Cheng, "Application status and prospect of big data and artificial intelligence in oil and gas field development," *Journal of China University of Petroleum*, vol. 44, issue 4, pp. 1-11, 2020.

[9] K. E. Dierckens *et al*., "A data science and engineering solution for fast k-means clustering of big data," *IEEE TrustCom*/*BigDataSE*/*ICESS*, pp. 925-932, 2017.

[10] C. K. Leung, "Data science for big data applications and services: Data lake management, data analytics and visualization," *Big Data Analyses, Services, and Smart Data*, pp. 28-44, 2021.

[11] O. R. H. G. Virginia *et al*., "Sentiment classification of film reviews using IB1," in *the Proc. the 7th International Conference on Intelligent Systems, Modelling and Simulation* (*ISMS*), 25-27 Jan. 2016, pp. 78-82, doi: 10.1109/ISMS.2016.38

[12] T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms," in *the Proc. International Conference on Microelectronics, Computing and Communications* (*MicroCom*), 23-25 Jan. 2016, pp. 1-6, doi: 10.1109/MicroCom.2016.7522583.

[13] V. Kotu and B. Deshpande, "Predictive analytics and data mining: Concepts and practice with RapidMiner," Morgan Kaufmann Publishers Inc, 2014.

[14] K. R. Dalal, "Review on application of machine learning algorithm for data science," in *Proc. 2018 3rd International Conference on Inventive Computation Technologies* (*ICICT*), Coimbatore, India, 2018, pp. 270-273, doi: 10.1109/ICICT43934.2018.9034256.

[15] R. Ahmed, M. Faizan, and A. I. Burney, "Process mining in data science: A literature review," in *Proc. 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics* (*MACS*), Karachi, Pakistan, 2019, pp. 1-9, doi: 10.1109/MACS48846.2019.9024806.

[16] G. Piatetsky. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. [Online]. Available: https://www.kdnuggets.com/2014/10/crisp-dm-topmethodology-analytics-data-mining-data-scienceprojects.html

[17] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc., 2000.

[18] SAS Institute, "Getting started with SAS enterprise miner 12.3," 2013.

[19] T. Kubiak and D. W. Benbow, "The certified six sigma black belt handbook," Milwaukee, WI: ASQQuality Press. 2005.

[20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[21] K. D. Kolo, S. A. Adepoju, and J. K. Alhassan, "A decision tree approach for predicting students academic performance," *International Journal Education and Management Engineering*, vol. 5, pp. 12–19, 2015.

[22] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student academic performance prediction model using decision tree and fuzzy genetic algorithm," *Procedia Technology*, vol. 25, pp. 326–332, 2016.

[23] A. B. Raut and M. A. Nichat, "Students performance prediction using decision tree," *Int. J. Comput. Intell. Res.*, vol. 13, no. 7, pp. 1735–1741, 2017.

[24] A. S. Olaniyi, S. Y. Kayode, H. M. Abiola, S. I. T. Tosin, and A. N. Babatunde, "Students performance analysis using decision tree algorithms," *Annals. Computer Science Series*, vol. 15, no. 1, 2017.

[25] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, "Student academic performance prediction by using decision tree algorithm," in *Proc. 4th International Conference on Computer and Information Sciences*, 2018.

[26] A. Zaldivar-Colado, J. A. Aguilar-Calderon, O. V. Garcia-Sanchez, C. E. Zurita-Cruz, M. M. Estrada, and R. Bernal-Guadiana, "Artificial neural networks for the prediction of students academic performance," in *Proc. 8th International Technology, Education and Development Conference*, 2014.

[27] H. T. Binh and B. T. Duy, "Predicting students' performance based on learning style by using artificial neural networks," in *Proc. 9th International Conference on Knowledge and Systems Engineering*, Vietnam, 2017.

[28] L. Gerritsen, "Predicting student performance with neural networks," Doctoral dissertation, Tilburg University, 2017.

[29] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata, "A neural network approach for students' performance prediction," in *Proc. the Seventh International Learning Analytics & Knowledge Conference*, 2017, pp. 598-599.

[30] D. P. Bendangnuksung, "Students performance prediction using deep neural network," *Int. J. Appl. Eng. Res.*, vol. 13, no. 2, pp. 1171–1176, 2018.

[31] H. Shaziya, R. Zaheer, and G. Kavitha, "Prediction of students performance in semester exams using a Naïve Bayes classifier," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 4, no. 10, 2015, pp. 9823–9829.

[32] M. Makhtar, H. Nawang, and S. N. W. Shamsuddin, "Analysis on students performance using Naïve Bayes classifier," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 16, 2017, pp. 3993–3999.

[33] V. Patil, S. Suryawanshi, M. Saner, V. Patil, and B. Sarode, "Student performance prediction using classification data mining techniques," *International Journal of Scientific Development and Research*, vol. 2, no. 6, 2017, pp. 163-167.

[34] F. Razaque, N. Soomro, S. A. Shaikh, S. Soomro, J. A. Samo, N. Kumar, and H. Dharejo, "Using naïve bayes algorithm to students' bachelor academic performances analysis," in *Proc. 2017 4th IEEE International Conference on*: *Engineering Technologies and Applied Sciences* (*ICETAS*), pp. 1-5.

[35] Y. Divyabharathi and P. Someswari, "A framework for student academic performance using naive Bayes classification technique," *J. of Advancement in Engineering and Technology*, vol. 6, no. 3, 2018, pp.1-4.

[36] M. G. Asogbon, O. W. Samuel, M. O. Omisore, and B. Ojokoh, "A multi-class support vector machine approach for students academic performance prediction," *International Journal of Multidiscplinary and Current Research*, vol. 4, 2016, pp. 210-215.

[37] G. Pratiyush and S. Manu, "Classifying educational data using support vector machines: A supervised data mining technique," *Indian J. Sci. Technol.*, vol. 9, no. 34, 2016.

[38] A. Kadambande, S. Thakur, A. Mohol, and A. M. Ingole, "Predicting students performance system," *International Research Journal of Engineering and Technology*, vol. 4, no. 5, 2017, pp. 2814-2816.

[39] S. A. Oloruntoba and J. L. Akinode, "Student academic performance prediction using support vector machine," *International Journal of Engineering Sciences and Research Technology*, vol. 6, no. 12, 2017, pp. 588-597.

[40] Y. Han and W. Lam, "Exploring query matrix for support pattern based classification learning," *Advances in Machine Learning and Cybernetics, Lecture Notes in Computer Science*, vol. 3930/2006, pp. 209-218.

[41] M. Zhanga and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, issue 7, July 2007, pp. 2038-2048.

[42] N. Ishii, Y. Hoki, Y. Okada, and Y. Bao, "Nearest neighbor classification by relearning," in *Proc. the 10th International Conference on Intelligent Data Engineering and Automated Learning* (*IDEAL'09*), 2009, pp, 42-49.

[43] B. Minaei-Bidgoli, D. A. Kashy, G. Kortmeyer, and W. F. Punch, "Predicting student performance: An application of data mining methods with an educational Web-based system," *33rd Annual Frontiers in Education*, vol. 1, 2003 pp. 2-18.

[44] Y. Zou, A. An, and X. Huang, "Evaluation and automatic selection of methods for handling missing data," *IEEE International Conference on Granular Computing*, vol. 2, 2005, pp. 728-733.

[45] M. M. Quadri and N. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global Journal of Computer Science and Technology*, vol. 10, no. 1.

[46] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," *Economic Review*, vol. 10, no. 1.

[47] S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions," *Expert Systems with Applications*, vol. 41, no. 14, 2014, pp. 6400–6407.

[48] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás-Martínez, "Data mining algorithms to classify students," *Educational Data Mining*, 2008.

[49] B. M. Bidgoli, D. Kashy, G. Kortemeyer, and W. Punch, "Predicting student performance: An application of data mining methods with the educational web-based system lon-capa," in *Proc. ASEE/IEEE Frontiers in Education Conference*, 2003.

[50] Z. Ibrahim and D. Rusli, "Predicting students academic performance: Comparing artificial neural network," *Decision Tree and Linear Regression, in*: *21st Annual SAS Malaysia Forum*, 5th September, 2007.

[51] K. Bunkar, U. K. Singh, B. Pandya, and R. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification," in *Proc. 2012 Ninth International Conference on Wireless and Optical Communications Networks* (*WOCN*), IEEE, 2012, pp. 1–5.

[52] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting student performance: a statistical and data mining approach," *International Journal of Computer Applications*, vol. 63, no. 8, 2013, pp. 35–39.

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

[54] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," in *Proc. 2014 IEEE International in*: *Advance Computing Conference* (*IACC*), IEEE, 2014, pp. 549-554.

[55] P. M. Arsad, N. Buniyamin, and J.-L. A. Manan, "A neural network students' performance prediction model (nnsppm)," in *Proc. 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications* (*ICSIMA*), IEEE, 2013, pp. 1–5.

[56] T. Wang and A. Mitrovic, "Using neural networks to predict student's performance," in *Proc. International Conference on Computers in Education*, IEEE, 2002, pp. 969–973.

[57] V. Oladokun, A. Adebanjo, and O. Charles-Owaba, "Predicting students academic performance using artificial neural network: A case study of an engineering course," *The Pacific Journal of Science and Technology*, vol. 9, no. 1, 2008, pp 72–79.

[58] D. M. S. A. Kumar, "Appraising the significance of self regulated learning in higher education using neural networks," *International Journal of Engineering Research and Development*, vol. 1, no. 1, 2012, pp. 9-5.

[59] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decision Analytics*, vol. 2, no. 1, 2015, pp. 1–25.

[60] B. M. Bidgoli, D. Kashy, G. Kortemeyer, and W. Punch, "Predicting student performance: An application of data mining methods with the educational web-based system lon-capa," in *Proc. ASEE/IEEE Frontiers in Education Conference*, 2003.

[61] W. Hamal and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems," *Intelligent Tutoring Systems*, Springer, 2006, pp. 525–534.

[62] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of student academic performance by an application of data mining techniques," in *Proc. International Conference on Management and Artificial Intelligence IPEDR*, vol. 6, 2011, pp. 110–114.

Copyright © 2022 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).

Kitsadaporn Jantakun is a lecturer in computer education, bachelor's degree program, Faculty of Education, Roi Et Rajabhat University, Roi Et, Thailand. Her research interests include data science, game-based learning, communities of practice, computer-mediated communication, creativity, knowledge management, and innovation.

Thiti Jantakun is an assistant professor in computer education with a bachelor's degree program in the Faculty of Education at Roi Et Rajabhat University, Roi Et, Thailand. His research interests include AI, big data, digital game-based learning, design thinking, and STEAM education.

Thada Jantakoon is an assistant professor in information and communication technology for education, Ph.D. Program, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand. His research interests include STEAM education, design thinking, data science, educational technology, and instructional technology.