of times students have accessed course materials within the designated time frame for each stage (e.g., Week 1 to Week 7 for Stage 1, Week 1 to Week 12 for Stage 2, and so on). This information is a proxy for student engagement and interaction with the course content, potentially offering early signs of potential challenges.

G. Data Modeling and Evaluation

During this stage, predictive models will be built using data engineered from earlier steps in data preprocessing and analysis. The models chosen for this process include a mix of machine learning methods, such as Decision Trees (DT), Support Vector Classifiers (SVC), and possibly more, depending on how they are applied. These models are then trained on a segmented dataset by splitting it into three parts: a training set for model training, a validation set for adjusting model settings and choosing the best model, and a test set for evaluating how well the models perform on data.

Every model is trained on a training set with specific settings and parameters tailored to its approach. After the training phase, the models' abilities are thoroughly tested using the validation set [24]. Important measures like accuracy, precision, recall, and F1-score are calculated to determine how well each model can predict student outcomes and identify students who might be at risk based on the features engineering. The selection of appropriate metrics is crucial for effectively evaluating the performance of a predictive model [24], [25]. This study utilizes a combination of commonly employed metrics to comprehensively assess the effectiveness of the developed models in predicting at-risk students.

1) Accuracy: Accuracy, a widely used metric, is calculated by dividing the number of correctly classified instances by the total number of samples in the dataset [6]. While intuitive, accuracy can be misleading in scenarios with imbalanced datasets, where the model might be biased towards the majority class [6]. The formula for accuracy is presented below (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where:
TP (True Positive) = Correctly predicted positive cases
TN (True Negative) = Correctly predicted negative cases
FP (False Positive) = Incorrectly predicted positive cases (Type I error)
FN (False Negative) = Incorrectly predicted negative cases (Type II error)

2) Precision: Precision focuses on the proportion of truly positive predictions [6]. It is calculated by dividing the number of correctly identified positive cases (TP) by the total number of positive predictions the model makes (TP + FP). The formula for precision is as follows (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3) Recall: Recall, also known as sensitivity, measures the model's ability to identify all actual positive cases [6]. It is calculated by dividing the number of correctly identified positive cases (TP) by the total number of actual positive cases in the dataset (TP + FN). The formula for recall is presented in (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4) F1-score: The F1-score provides a harmonic mean of precision and recall, offering a more balanced view of the model's performance [6]. It is particularly valuable in imbalanced datasets, where solely relying on accuracy can be misleading. The F1-score is calculated using the following formula (4).

$$F1-score = 2x \frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

III. Results and Discussion

The strategy of train-test-validation split was employed to ensure a balanced and unbiased representation of both student categories ("at-risk" and "pass") across the training, validation, and test datasets. The training set is 70% and the test set is 30%. The validation is using the same size of data. This approach is crucial for preventing bias that could occur if certain student categories were disproportionately represented in specific datasets. By maintaining a balanced distribution, the evaluation of the model's performance on unseen data becomes more reliable and applicable.

The Decision Tree Classifier, set at a maximum depth of 3, achieved high accuracy as in Table II across all datasets: 0.9911 on the training set, 0.9770 on the validation set, and 0.9744 on the test set. This model achieved a precision of 1.00 and a recall of 0.58, resulting in an F1-score of 0.74. The validation set showed improved precision at 0.91 and recall at 0.83, with an F1-score of 0.87. The test set exhibited a balanced performance with a precision of 0.92 and a recall of 0.92, leading to an F1-score of 0.92. This result indicates outstanding performance but also raises concerns about the possibility of overfitting. In this case, the model might have become too proficient in memorizing the training data, possibly including irrelevant or noisy patterns. This could lead to overly positive performance metrics that might not hold up when applied to new, unseen data.

The Support Vector Classifier (SVC) demonstrated a good balance between performance on the training and validation sets, with accuracy scores of 0.9947 and 0.9862, respectively. The test accuracy of 0.9764 further supports this, suggesting that SVC can learn from the training data while still being able to generalize well to new, unseen data in the validation set. The test accuracy of 0.9655 further supports this idea. The performance indicators of the SVC were notably impressive, boasting a precision of 1.00 and a recall rate of 0.94 on the training data, leading to an F1-score of 0.97. When applied to the validation data, it managed a precision of 0.98 and a recall rate 1.00, achieving an F1-score of 0.99. The performance on the test data mirrored this consistency, with precision and recall reaching 1.00 and an F1-score of 1.00. This uniformity underscores this model's ability to accurately pinpoint the best hyperplane that maximizes the separation between the two student categories ("at-risk" and "pass"), rendering it a strong predictor of student outcomes.

Gaussian Naive Bayes, a classifier based on the assumption that features are independent within each class, performed exceptionally well across all datasets. The model achieved accuracy scores of 0.9994 on the training set, 0.9908 on the validation set, and 0.9902 on the test set. The evaluation metrics demonstrated exceptional accuracy and the ability to

correctly identify instances, with the training data achieving a precision of 1.00 and a recall of 0.94, leading to an F1 score of 0.97. The evaluation data also displayed an accuracy of 0.99 and recall of 1.00, with an F1-score of 0.99. The testing data continued to perform at this level, with precision and recall at 1.00 and an F1-score of 1.00. These outcomes indicate the model's success in categorizing student information and its robustness in applying it to new data. However, it's essential to consider the assumption that features are independent, as this may not apply universally to all datasets, which could impact the model's performance across various situations.

Logistic Regression, designed for binary classification tasks, was used to predict the probability of a student being at risk based on their features [26], [27], [28]. The model provided probability scores for each class, and its accuracy was compared with the other models. Logistic Regression achieved training, validation, and test accuracies of 0.9397, 0.9401, and 0.9350, respectively. In the training data, the model achieved a precision of 0.48 and a recall of 1.00, resulting in an F1 score of 0.65. The data used for validation demonstrated an accuracy of 1.00 and recall of 0.94, with an F1-score of 0.97. The data from the test set also showed consistent performance, with both precision and recall values at 0.92, leading to an F1-score of 0.92. These findings indicate that although Logistic Regression was effective, it fell short of fully understanding the intricate connections between student characteristics and their risk levels compared to models like SVC or Gaussian Naive Bayes.

TABLE II
COMPARISON OF ACCURACY BETWEEN EACH MODEL

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Decision Tree	0.9911	0.9770	0.9744
Support Vector Classifier (SVC)	0.9947	0.9862	0.9764
Gaussian Naive Bayes	0.9994	0.9908	0.9902
Logistic Regression	0.9397	0.9401	0.9350

After conducting a detailed comparison in Table II and choosing the best model for predicting at-risk students, valuable insights were gained by looking at how well each model did on training, validation, and test data. The Decision Tree showed impressive accuracy during these evaluations, which made fitting the data too closely and looked further into its ability to work well with new data. On the other hand, the Support Vector Classifier (SVC) came out as a strong contender, performing consistently well across all datasets and proving itself good at making predictions.

The Gaussian Naive Bayes model did remarkably well, especially because it matched the idea that each class has independent features. However, it's important to remember how true this is and how well it works in situations where this independence might not be accurate. Similarly, Logistic Regression did a good job. Still, its lower accuracy levels suggest it might struggle to understand the full depth of the relationships more than some other models in our study.

Considering these results, the SVC model is the best option for predicting at-risk students in this situation because it performs well overall and finds the best line for separating classes. Still, there is a need to explore the other models more, especially on preventing Decision Tree overfitting and looking more closely at how the Gaussian Naive Bayes model treats features independently in each class to improve how well it works in different situations. Thus, future research could greatly benefit from a closer look at these aspects to make better choices for predicting outcomes in educational settings.

The study by [8] used a traditional approach, relying on Excel documents to analyze data instead of a data lake framework. In [8], it was found that early prediction models were viable when certain performance indicators like accuracy, precision, recall, and F1 scores were between 75% and 85%. Conversely, this study achieved impressive outcomes with the Gradient Boosting model, showing varying performance levels throughout the course. When predicted using complete course data, the GB model reached its peak performance in precision, recall, and F1 scores (91.79%, 98.48%, and 94.84%, respectively). However, the model's performance dropped when predicted from earlier course stages, such as W1—W14, experiencing a precision score decrease of 33.1%. The SVC model also performed well, with precision scores of 90.35%, recall scores of 96.93%, and F1-scores of 93.55% for the full course length, though it saw a decrease in performance in earlier stages. Despite these drops, the W1—W14 stage was the closest to meeting the performance benchmarks set by [6], achieving a recall score of 83.5%. This underscores the crucial role of final exam grades in predicting outcomes and the benefits of employing a data lake framework for more comprehensive data analysis and early intervention strategies.

IV. CONCLUSION

This study highlights the significant obstacles Universiti Putra Malaysia (UPM) encounters in effectively managing and analyzing educational data. The disjointed and inconsistent spread of data across various university systems [29], [30], such as the Student Information System (SIS) and the Learning Management System (PutraBlast), severely limits UPM's capacity to obtain a full understanding of student achievement. This absence of a unified data storage system adds complexity, making it difficult for the university to extract valuable insights and provide timely support to students who might be struggling. Furthermore, the scattered nature of educational data complicates efforts to ensure data security and privacy, as confidential student information is spread across different platforms, heightening the risk of unauthorized access and data leaks. This situation threatens the accuracy of student records and raises questions about adherence to data protection laws.

A solution has been proposed to tackle these pressing issues. This solution consolidates various educational data sources into a unified data framework using advanced data management technologies, especially by utilizing a data lake environment supported by Dremio. By adopting this strategy, the data ingestion and transformation processes will become more efficient, significantly improving data management within the university. This, in turn, will facilitate a more precise predictive analysis of student success, ultimately benefiting the entire UPM community.

Adopting this suggested solution will not only help UPM address its current challenges in managing educational data but will also set the university up for future success in using data-driven insights to improve student outcomes and overall academic quality. By adopting modern data management technologies, UPM can ensure its leadership in innovation and educational excellence in Malaysia and beyond. Recognizing the impact of data on student performance and the overall quality of education is crucial for creating an environment that supports ongoing improvement and personalized learning experiences. By understanding the obstacles faced by universities like UPM and the proposed solutions, students can better appreciate the efforts to enhance their educational journey and contribute to a more data-aware educational environment.

This study has laid a solid groundwork for using data architecture to improve students' performance. However, there are many exciting areas for further study and innovation. A major focus is making the most of the data lake by using advanced analytics methods. Adding complex machine learning algorithms and AI models could significantly enhance the system's ability to predict student success [7]. By exploring detailed patterns in the data, researchers can create models that accurately identify students who might struggle early on, allowing for prompt support and customized help.

Developing effective data visualizations and user-friendly dashboards is essential for sharing insights from the data with various stakeholders, including educators, administrators, and students. Creating interactive visual tools helps in understanding trends in student performance and identifying strengths and areas needing improvement, which supports making decisions based on data. Ensuring its privacy, accuracy, and safety is a top priority when managing and securing student data. Implementing strict data management policies and security measures is key to protecting sensitive information. Regular data security checks and reviews are necessary to meet relevant standards and build trust within the university community.

Growing UPM's data collection by incorporating various sources can provide a more comprehensive and detailed view of student achievement. By combining information from systems like student feedback, alumni surveys, or external evaluations, the analysis can be more thorough, leading to more accurate predictions and useful insights. By focusing on these important areas, UPM is well-positioned to fully leverage data in making decisions, ultimately creating a more personalized and effective educational experience for its students.

ACKNOWLEDGMENT

The Faculty of Computer Science and Information Technology, University Putra Malaysia, funded this research.

REFERENCES

[1] H. Jahankhani, A. Jamal, G. Brown, E. Sainidis, R. Fong, and U. J. Butt, Eds., AI, Blockchain and Self-Sovereign Identity in Higher Education. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-33627-0.

[2] R. Raju, R. Mital, and D. Finkelsztein, "Data Lake Architecture for Air Traffic Management," *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, pp. 1–6, Sep. 2018, doi:10.1109/dasc.2018.8569361.

[3] P. Wieder and H. Nolte, "Toward data lakes as central building blocks for data management and analysis," *Frontiers in Big Data*, vol. 5, Aug. 2022, doi: 10.3389/fdata.2022.945720.

[4] A. Cuzzocrea, "Big Data Lakes: Models, Frameworks, and Techniques," *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–4, Jan. 2021, doi:10.1109/bigcomp51126.2021.00010.

[5] D. Martinez-Mosquera, V. Beltran, D. Riofrio-Luzcando, and J. Carrion-Jumbo, "Data Lake Management for Educational Analysis," *2022 IEEE Sixth Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–5, Oct. 2022, doi: 10.1109/etcm56276.2022.9935751.

[6] S. M. M. Muin et al., "Predicting academic student performance based on e-learning platform engagement using learning management system data," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, pp. 1859–1866, 2023. doi: 10.17762/ijritcc.v11i9.9178.

[7] H. S. Brdesee, W. Alsaggaf, N. Aljohani, and S.-U. Hassan, "Predictive Model Using a Machine Learning Approach for Enhancing the Retention Rate of Students At-Risk," *International Journal on Semantic Web and Information Systems*, vol. 18, no. 1, pp. 1–21, Mar. 2022, doi: 10.4018/ijswis.299859.

[8] C. Manco, T. Dolci, F. Azzalini, E. Barbierato, M. Gribaudo and L. Tanca, "HEALER: A data lake architecture for healthcare", *Proc. Workshops EDBT/ICDT Joint Conf.*, pp. 1-10, 2023, Mar. 19, 2024, [online] Available: https://ceur-ws.org/Vol-3379/DataPlat_2023_602.pdf.

[9] G. Weintraub, E. Gudes, S. Dolev, and J. D. Ullman, "Optimizing Cloud Data Lake Queries With a Balanced Coverage Plan," *IEEE Transactions on Cloud Computing*, vol. 12, no. 1, pp. 84–99, Jan. 2024, doi: 10.1109/tcc.2023.3339208.

[10] D. Mazumdar, J. Hughes, and J. B. Onofre, "The Data Lakehouse: Data Warehousing and More," 2023. [Online]. Available: arXiv:2310.08697. [Accessed: Mar. 19, 2024].

[11] F. Qiu et al., "Predicting students' performance in e-learning using learning process and behaviour data," *Scientific Reports*, vol. 12, no. 1, Jan. 2022, doi: 10.1038/s41598-021-03867-8.

[12] J. Fan, "A big data and neural networks driven approach to design students management system," *Soft Computing*, vol. 28, no. 2, pp. 1255–1276, Dec. 2023, doi: 10.1007/s00500-023-09524-8.

[13] R. P. d. C. C. Macedo, "Implementation of a Data Lake in a Microservices Architecture," Master dissertation, Department Information, University of Lisbon, Portugal, 2024, [Online] Available: http://hdl.handle.net/10451/63925.

[14] R. Asokan, D. P. Ruiz, and S. Piramuthu, Eds., Smart Data Intelligence. Springer Nature Singapore, 2024. doi: 10.1007/978-981-97-3191-6.

[15] "What is a non-relational database?" (n.d.). [Online]. Available: https://www.mongodb.com/resources/basics/databases/non-relational. Accessed: Mar. 19, 2024.

[16] "What is a cloud data lake?" (n.d.). [Online]. Available: https://www.dremio.com/resources/guides/cloud-data-lakes/. Accessed: Sept. 24, 2024.

[17] C. Cuello, "Data Ingestion vs. Data Integration: Know the differences for efficient data management," Dec. 5, 2023. [Online]. Available: https://rivery.io/data-learning-center/data-ingestion-vs-data-integration/. [Accessed: Mar. 19, 2024].

[18] "Data Ingestion vs. Data Integration: What Sets Them Apart?" Feb. 27, 2024. [Online]. Available: https://airbyte.com/data-engineering-resources/data-ingestion-vs-data-integration. Accessed: Mar. 19, 2024.

[19] M. Garcia, "The Evolution of Data Pipelines: ETL, ELT, and the Rise of Reverse ETL," CORE, Oct. 2, 2023. [Online]. Available: https://dzone.com/articles/the-evolution-of-data-pipelines.

[20] "What is Data Transformation?" (n.d.). [Online]. Available: https://www.tibco.com/glossary/what-is-data-transformation#, Accessed: Mar. 19, 2024.

[21] D. Ushasree, A. V. Praveen Krishna, and Ch. Mallikarjuna Rao, "Enhanced stroke prediction using stacking methodology (ESPESM) in intelligent sensors for aiding preemptive clinical diagnosis of brain stroke," *Measurement: Sensors*, vol. 33, p. 101108, Jun. 2024, doi:10.1016/j.measen.2024.101108.

[22] J. Xiong et al., "Deep Learning-Based Open Source Toolkit for Eosinophil Detection in Pediatric Eosinophilic Esophagitis," Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.06333.

[23] G. Siemens, "Learning Analytics," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, Aug. 2013, doi:10.1177/0002764213498851.

[24] "Streamlining Predictive Analytics with Scikit-Learn," (n.d.). [Online]. Available: https://www.osedea.com/insight/streamlining-predictive-analytics-with-scikit-learn. Accessed: Mar. 19, 2024.
[25] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)," *Education and Information Technologies*, vol. 28, no. 7, pp. 8299–8333, Dec. 2022, doi: 10.1007/s10639-022-11536-0.
[26] "Model Performance," (n.d.). [Online]. Available: https://fastercapital.com/keyword/model-performance.html. Accessed: Sept. 24, 2024.
[27] Z. Liu, W. Chen, C. Liu, R. Yan, and M. Zhang, "A data mining-then-predict method for proactive maritime traffic management by machine learning," *Engineering Applications of Artificial Intelligence*, vol. 135, p. 108696, Sep. 2024, doi: 10.1016/j.engappai.2024.108696.
[28] D. T. Larose, "Data Mining Methods and Models," John Wiley & Sons, Inc, Nov. 2005, doi: 10.1002/0471756482.
[29] P. Pooja and R. Bhalla, "A Review Paper on the Role of Sentiment Analysis in Quality Education," *SN Computer Science*, vol. 3, no. 6, Sep. 2022, doi: 10.1007/s42979-022-01366-9.
[30] P. Rangnekar, "What is Mobility in App Development? Key Insights!," Jan. 2024. [Online]. Available: https://www.biz4solutions.com/blog/category/uncategorized/page/13/. Accessed: Mar. 19, 2024.