ŌTAUTAHI CHRISTCHURCH AOTEAROA NEW ZEALAND
35TH AUSTRALASIAN ASSOCIATION FOR ENGINEERING EDUCATION ANNUAL CONFERENCE
8-11 DEC 2024
THE ENGINEER AND THE WORLD

Leveraging Machine Learning for Insights into Engineering Course Design

Shelby Nuttall, Michael Crocco, and James Salamy
Monash University
Corresponding Author Email: james.salamy@monash.edu

CONTEXT

The impacts of subject study order on student success are nuanced, difficult to measure, and inherently interconnected to a number of other factors. Universities typically use intuitive or crude statistical approaches to judge the effectiveness of course “layouts” and any changes to them.

As a diversity of degree offerings becomes the norm, it is imperative to establish flexibility in course progression so that capable students are not unjustifiably delayed or discouraged in their studies. Tools that can capture the importance of certain course properties are crucial in the validation of both traditional inputs and novel course layouts that empower students.

PURPOSE

This study aims to produce a quantitative framework to validate qualitative human observations regarding course progression and student success. Drawing from an hypothesis, the framework should provide insight into the relative significance of course layout choices, and their impact on performance. This framework may also be applied to verification on an individual basis, to support student advisors in recommending the most effective pathway to mitigate the effect of a failure on a student’s ultimate goals.

APPROACH

This work develops a tool that uses a simple Machine Learning (ML) model to provide insight, via interpretability techniques into the relative importance of different aspects of prior study on student performance in a target subject at an aggregate (cohort) level. This tool can then be used to validate human driven hypotheses and insights.

The tool ingests processed numerical data drawn from a student’s prior studies at university. This data is broken down into three indicators calculated for all of prior studies - grade category, time since attempt, and attempts required to pass.

ACTUAL OUTCOMES

A selection of mechanical engineering and electrical engineering subjects were selected for analysis. Neural network models were able to consistently identify fail grades with an average accuracy of 93.6%. The model is effective at identifying subjects that contain relevant skills or knowledge beyond the ‘direct’ requirements defined by official prerequisites. In many cases, these subjects were stronger predictors of performance than prerequisite subjects.

CONCLUSIONS

This study shows the potential of ML techniques to assist academics in identifying significant factors that influence student success. The models generated are suitable to incorporate into program design processes to identify opportunities imperceptible to traditional analyses.

KEYWORDS

course structuring; academic progression; data analysis; machine learning.

Introduction

This paper focuses on an implementation of Machine Learning (ML) Neural Networks (NNs) to estimate the impacts of prior educational outcomes on the success or failure of a subsequent subject. In an engineering course with significant freedom of study sequence, optimal pathways could be identified based on past student results. This approach may allow an approach to course (program) design which focuses on threshold concepts, heretofore unidentifiable, for study progression, rather than a traditional focus on prerequisite content.

For example, when asking whether a student should be allowed to study Fluid Dynamics II without having successfully completed Fluid Dynamics I, course designers ought, perhaps, consider more than the two-step progression implied. Instead, appropriate analysis may point to other skills, outside of the field of Fluid Dynamics that are true precursors to success in this subject.

While many ML approaches have been implemented to predict student success, this study intends to prove a concept which could be implemented in the redesign of a course as well as the evaluation of any changes made.

Ethical approval for this study was granted by the Monash University Human Research Ethics Committee (MUHREC) with approval number 38711.

Background

Machine Learning can be used to predict behaviour in many, varied domains (Hohman et al., 2017), and offers improved predictive performance when compared to logistic regression (Senders et al., 2018). It is recognised that education and its associated field of learning analytics have potential benefit in the application of these methods with the aim of improving educational outcomes (Han et al., 2024).

ML techniques of various types (Balaji et al., 2024) have been applied to predict student performance in educational settings. Random Forest models are judged by Han et al. (2024) to demonstrate the lowest standard error and highest correctness among a group of models which includes Logistic Regression, Random Forest, K-Nearest Neighbours, Support Vector Regression, and Gaussian Naive Bayes when forecasting course results (learning outcomes in a subject, based on learner data from various sources).

Cumulative grade point average is, unsurprisingly, the most frequently used attribute when predicting educational outcomes (Rahul & Katarya, 2023). This equates to a statement that strong students will continue to be strong in the future, while weaker students will remain relatively weak; also a general finding of Asif et al. (2017). While such analyses are useful when informing student interventions or predicting student dropout (Bottcher et al., 2020), they are less constructive when informing the design of a course (program of study). Other approaches use psychological and behavioural data (Saha et al., 2023) or other demographic and study habit attributes (Limanto et al., 2023) to predict student performance, but with similar shortcomings.

At the same time, “the application of…machine learning techniques in educational mining [(learning analytics)] is still limited” (Balaji et al., 2024), primarily to this aim of informing individual student predictions (Senders et al., 2018; Sahiri et al., 2015). While useful in the classroom, larger scale course modifications are not enabled by the various, established approaches.

Method

Where established approaches are inappropriate for making large scale observations about a course, we propose a class of small, cost efficient models that can be used together to inform course design validity and potential academic risk factors. Contrary to traditional statistics approaches, using machine learning provides a mechanism for the analysis of large, complex multivariate datasets without requiring expert knowledge.

Given the multitude of possible subject orderings and course combinations common at Monash University, it would be a highly difficult task to attempt to create a single general model that is able to capture all of these interactions within an acceptable computational budget. A simple neural network template is instead used to facilitate the training of a family of similar models for single subject prediction on a subset of data relevant to those subjects. Once such models are constructed and trained, we utilise a suite of result visualization and analysis tools to compare and interpret results between subject models.

Data Collection

Data is collected from university databases and anonymised by ID hashing, and the deletion of other identifying fields before storage, with the hash key securely held by the chief investigator. The data entails an entry for every subject attempt by every student enrolled in any Faculty of Engineering subject over a period of interest. Entries include subject result, grade, mark, date and the unique (hashed) student identifier to build training examples across multiple subjects.

Data Filtering and Preparation

In order to generate training examples, the dataset is filtered to extract a subset of factors into an embedding that captures key information about a student's studies before attempting a subject of interest. For each student-subject pair, we identify 3 key factors: **Mark** - representing bins of performance categorised by the numerical mark (0-100), **Time**, representing how long ago (from the point of view of the subject of interest) a student passed the prior subject, and **Attempt**, which is used to identify students who have not yet attempted a subject (eg if a student has made a different progression decision, or it is an elective subject), and also as a metric for personal challenge, i.e. the number of attempts it took the student to pass the subject.

Machine learning techniques have a tendency to overfit to a dataset when the available information is sparse (Dabrian et al., 2023). As a result of the wide range of electives and courses offered to students, there are approximately 5000 individual subjects that have been taken by engineering students during the period of our dataset with almost 4000 of these being taken by fewer than 100 students across the entire available dataset. To prevent the proposed models from overfitting to these rare occurrences, the dataset was trimmed to include only core subjects for engineering specialisations as well as common first year subjects and popular electives.

Further filtering is applied to the raw mark field for each subject to generate 'bins', each representing a range of grades. This reduces sparsity as the number of grade categories (ranges) to decide between is strictly limited. Categorisation also presents the possibility of performing hyperparameter sweeps over boundary values to identify underlying clusters of students in different contexts. A larger and more complex model could potentially be used to predict a numerical output, but this would require additional compute resources relative to the model employed and result in less useful predictions.

The raw dataset also contains many flags around academic progress, for example, various categories of withdrawn results which provide context to a grade of 0. These entries, and other similar edge cases, are removed as they do not significantly contribute to course structure insights. This choice allows us to simplify the representation of a '0' mark in the embedding vector, corresponding to only a non-attempt (if Attempt is 0) or an actual mark of 0.

We propose that this feature selection is sufficient to represent a student's performance in a given subject. Mark is used to represent level of knowledge attainment at the end of the subject, Time represents the potential decay in familiarity with the material as time passes, and Attempts proposes to reflect difficulty with the subject (previous failure), and alternate course choices.

Data Transformation

Given that we want to generate models for many different subjects, the role a subject plays in the dataset can be a predictor of another result, or it can be the 'ground truth' used to train its own model. We call this second role the predictee subject.

As each model is designed to make predictions regarding a single subject (the predictee), the data must be transformed to the perspective of that subject. This requires the transformation of each of the time entries to represent the number of semesters that have elapsed between the completion of the corresponding predictor subject and the predictee, as shown in the equation below:

$$Elapsed\ time\ =\ Predictee\ time\ -Predictor\ time$$

Once this time value has been adjusted, any predictor subjects taken by a student after, or at the same time as, the predictee (which are identified by a 0 or negative values for relative, elapsed time) are removed from the embedding to avoid creating an anti-causal system which would not be of use for making predictions on new data. The mark value for the predictee subject corresponding to the expected grade category is also separated from the embedding.

Before using this processed data to train a model, it is important to apply some kind of balancing to the dataset. Machine learning algorithms, when provided with an uneven dataset (significantly more of one class of data entry than others), tend to favour the majority class (Krawczyk, 2016). Regarding overall academic performance in a subject, fail grades are typically a relatively small percentage of overall results, and most subjects have some form of normal distribution that leads to imbalance in the broad category ranges.

To mitigate the impact of data imbalance, the SMOTE data-augmentation technique is applied alongside over-sampling to synthetically generate plausible samples for underrepresented classes, resulting in a more even proportion of samples. SMOTE uses multi-dimensional interpolation to generate synthetic but representative samples for these datasets. This technique is superior to simple oversampling as it reduces the risk of overfitting to repeated single data points. Note, SMOTE is agnostic of any meaning assigned to inputs or relationships between them, as such the interpolated samples lack the practical meaning that the original samples have.

Model Design

The architecture of the proposed network determines how accurately the model may be able to perform as well as its training and analysis cost. Initially a simple neural network consisting of only linear layers was used. To better capture the relationship between groups of three key points (mark, attempt and time), a convolutional layer was introduced. Convolutional layers are generally used in image processing to identify significant types of features from groups of pixels. When applied to our dataset, the convolutional layer groups together the three elements of each subject and is able to learn an appropriate combination function to build a single numerical representation of 'knowledge' for that subject.

As our application requires models that are sufficiently low cost to be able to be retrained repeatedly across many different subjects and datasets, the final model architecture consists of only 4 layers (3 linear, 1 convolution) separated by ReLU and Dropout layers to reduce overfitting and increase the probability of convergence.

Before training the model, the dataset is split into three sections: a training set that will be used to inform the model's parameters, a validation set to check the model's progress during training (to avoid overfitting) and a test set that once the model has been trained, can be used to verify the accuracy of the model with data that has not be utilised during the training process.

The model is trained on a standard proportion of 72% of the available data, with a batch size of 10 using cross entropy loss to compare predicted and truth values. It is validated between each training step with 8% of the data. Once the model has been trained on the train and validate

datasets, the previously set aside data (20%) is run through the model and compared with the known truth values as a final check that the model has not overfitted to its training data.

Result Interpretation

Neural networks with their hidden layers and complex decision paths are famously lacking in interpretability (Zhuoyang & Feng, 2023). To allow useful observations to be made using these models, we choose to apply two primary toolsets: standard metrics and Shapley Value Analysis. The standard metrics report gives information on: a) model precision, which highlights how many of the predictions a model makes of a class are correct; b) recall, which provides an indicator of how many elements of a class were correctly predicted, and; c) an F1 score, which combines both precision and recall to represent general predictive performance per class in the dataset. The F1 score is used to measure the performance of the model, and highlight categories that the model struggles to predict.

Once the model has been verified with the above statistics, we can apply Shapley value analysis to determine the relative importance of different input factors on a model's predictions on an individual class scale (Shapley, 1953). This technique, initially proposed for applications in game theory, iteratively removes features from a section of data, measures the change in output prediction and uses this to inform the impact of each feature on a model's output (Merrick and Taly, 2020). This will allow a user to draw approximate conclusions regarding which subject performances have a significant impact on performance in a predictee subject as well as identifying through comparison which factors may have a larger than average impact on the distinction between failure and passing.

Results and Discussion

Model Analysis

A first consideration is the computation complexity of this model. This simple model, containing only 8700 parameters, can be effectively trained on a regular personal computer without the need for expensive GPU accelerators. The models presented in this paper were generated using a 2020 MacBook with an i5 quad-core processor, taking an average wall clock time of 3 minutes, and a worst case time of under 6 minutes to train. All models converged, although 2 models needed to be reinitialised and retrained in order to converge. This is not a major issue, as the computation cost at this phase is minimal.

Unusually for ML processes, the inference for the interpretation phase is significantly more expensive than the training phase, taking around 2 hours on the same hardware to simulate the interactions of features in the data drawn from a sample of 800 students. This phase could be significantly improved by the parallelisation of the implementation and the usage of a GPU.

Selection of category boundaries (bins) is a key determinant of performance, and represents the hyperparameters of this system. Initial experiments using university grade boundaries did not lead to satisfactory accuracy, leading to the reduced categorisation presented here. Breaking a student's performance into "Failure", "Pass", and "Good" categories signifies the tradeoff between providing enough depth to allow for course design decisions without sacrificing accuracy.

Figure 1 presents a sweep over several choices. The variance in performance across slight differences in the edge cases is evidence of the underlying complex groupings or clusters in the data, which vary between types of subjects, areas of study, and year levels. Most reasonable choices produce good results, with the optimal choice depending on the specific subject. Performing a binary classification (Pass or Failure) performs very strongly, and has potential to be used to identify common failure pathways directly, but provides less input on the majority of students who do pass, and is less useful in the context of generating course structure insights.

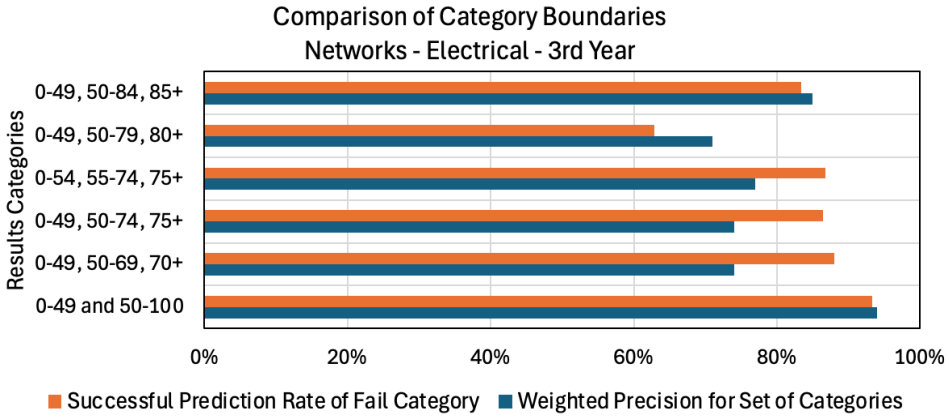

Figure 1: Sweep over a selection of boundary conditions for a sample subject from electrical engineering. These results demonstrate the rate of failing students each model successfully predicts, and the overall correctness of each model's predictions, weighted by the relative number of students in each category of the test set.

While analysis of individual categories and weighted accuracy is helpful for comparison across models and settings, it obfuscates a key piece of information evident in the confusion matrix presented in Figure 2 for the same subject as in Figure 1. Confusion matrices map the rate at which the model predicts each student category against the true category that student should have obtained. This reveals that this model (common to all subjects examined) is not 'equally' wrong when it selects the wrong category - it tends 'miss by one' category in most cases, and also tends to bifurcate the data between 'failing' and both 'passing' categories, to a extent that varies on the subject, but correctly predicts fail grades at an average rate 93.6%.

This view must also be considered to determine the best general categorisation for the group of subjects under consideration. For this paper, we have analysed a group consisting of all unique Mechanical and Electrical engineering core subjects across 3rd and 4th year, which means we are considering several potentially distinct underlying clustering patterns. We have therefore not tried to optimise the category borders for any one subject or area, and have adopted a relatively simple mapping of 0 to 49 as representing 'Failing', 50-69 as representing 'Passing' a subject, and 70 or above as achieving a 'Good' level of understanding, which are drawn from combining Monash University's grade boundaries.

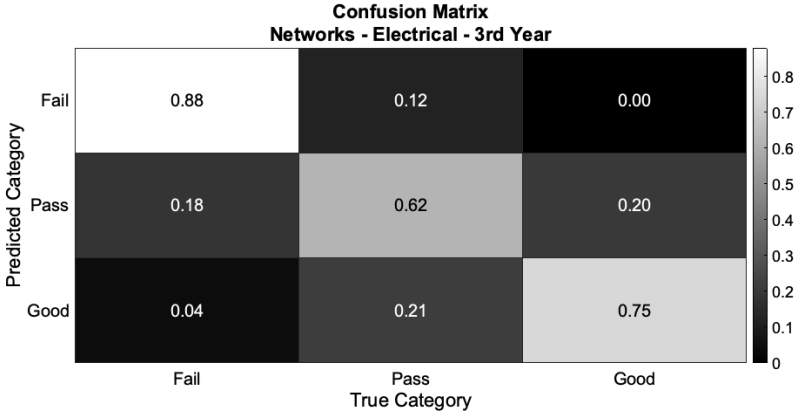

Figure 2: Confusion Matrix demonstrating the distribution of incorrect guesses for each class. This analysis has been performed on the model for the same subject presented in Figure 1, using the boundaries Fail: 0-49, Pass: 50-69, Good: 70-100 boundaries.

The model family is able to generalise and provide accuracy significantly better than random prediction over the range of subjects examined in this study, achieving similar performance in terms of generalised precision, as shown in Figure 3. This general applicability to engineering results without fine-tuning gives confidence to the following analysis on model interpretation.

Interpreting Subject Relationships

Using the Shapley technique, we observe the relative strength of the contribution of each factor. We present a 3rd year electrical engineering subject covering networking topics in Figure 4 as an example. This subject was chosen as an interesting analysis case for a number of reasons: it is a foundational subject that unlocks many electives, the authors are familiar with the subject, allowing us to act as subject matter experts, and there are no specific prerequisites, leading to a broad range of student experiences around prior subjects and placement with their degrees.

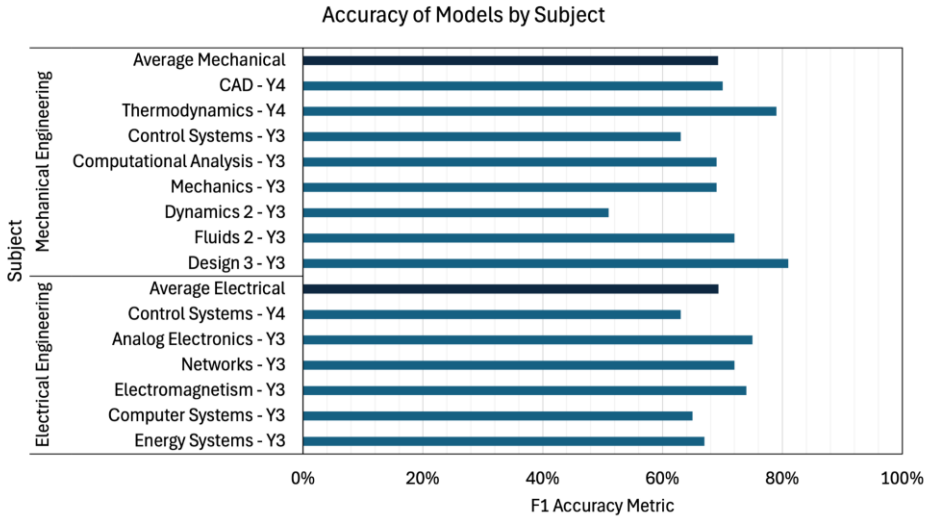

Figure 3: Comparison of model performance by subject across two disciplines. Discipline averages are also presented, with an overall average of 72% achieved across all models.

The percentage of blue weighting in each bar of Figure 4 demonstrates one of the consistent observations around the performance of these models, which is that they spend a disproportionate amount of effort on the classification of failure, which is also reflected in Figure 2. This is a reasonable priority of a tool to identify failings within a course structure, but limits the accuracy of the model in differentiating higher performing students more clearly. A potential future refinement is to separate the training into a pass-fail model and a specific grade model, which should significantly improve performance.

Figure 4 demonstrates the importance of marks in predictions, as 57% of the factors identified are grade categories. This demonstrates that categorised marks are a reasonable predictor of future performance, and there is a correlation between relevance of material and prediction strength. Another strong correlation is with similar assessment styles, such as a second year digital, or a structures & design subject in first year, correlating to the open project in this subject. This demonstrates that subjects that teach ‘engineering thinking’ in other contexts clearly impact performance later in the degree. This was not an obvious observation to us a priori.

In general, timing of subjects is not a significant factor, but the timing of key subjects, such as the common first year maths, alludes to patterns in progression (i.e. foundation subject requirements result in different orders, leading to different levels of success). This is an example of a potential gap in support that needs further investigation and potentially different advice or resources. The timing of the core 2nd year circuits most likely correlates to the cohort who are in double degrees, underloading and/or have failed at least one core subject, who may be on different paths that again needs unique advice. The lack of time factors also implies that the expected ‘degradation of skills with time’ effect is relatively unimportant in predicting performance later in the degree.

The presence of a significant subject attempts factor, such as 2nd year signals, corresponds to a subject that is the most similar in terms of material to the 3rd year networks subject. While performance in this subject is a more significant factor, the number of attempts captures students who have either completed it prior to networks (0), passed first time (1), or who needed more than one attempt (2+), which we would expect to significantly impact a prediction.

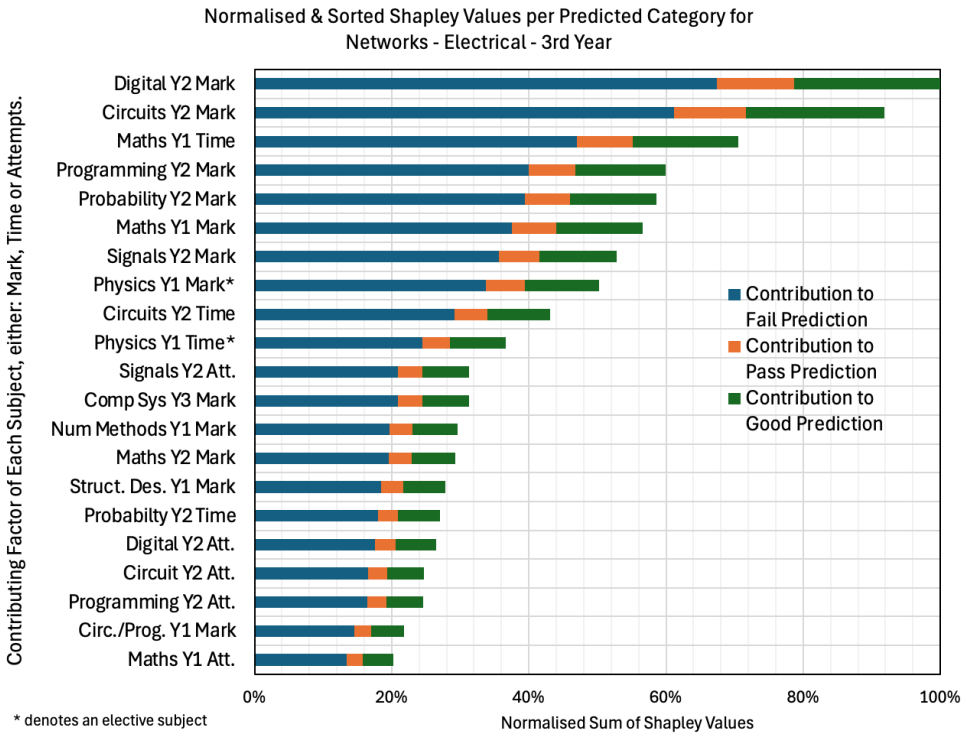

Figure 4: Shapley values for a given subject of interest, the same subject as in Figures 1 and 2. Values have been normalised and sorted such that the strongest contributor to an average (from a sample of 800) student's result is at the top, with the relative importance of each subject decreasing down the y axis. Only factors >20% as strong as the top rank are included for clarity.

Note that in many cases, pre-requisite subjects, though in the top 20 most significant features, are not the most predictive of outcomes in a subject. This suggests that mandated study progressions, in general, may not be as critical to success as has traditionally been held. In this specific case, a single elective subject reached the top 20 significant features threshold, and this was seen in other subject tests. This indicates that the list of core and non-core subjects analysed should be tailored specifically to the subject being modelled.

Conclusion

Neural networks of the type studied here show potential, with further refinement, to predict individual subject outcomes and allow the evaluation of study program pathways. Observations made from this analysis can contribute to and inform discussions with subject matter experts and program coordinators to guide optimisations and refinements as part of the course development process, where current techniques are largely qualitative. Using these models could potentially inform the validity of prerequisite listings as well as credit point requirements of individual subjects. Used in conjunction with qualitative evaluation, this approach could facilitate more effective course design.

Additionally, these low-cost models could be used to provide study planning support to individual students, allowing them to find optimal ordering for a set of subjects, or to assess the academic impact of overloading or prerequisite waiving to reduce the guesswork of these processes. Future extensions could include additional data, such as previous withdrawn attempts, previous educational data (VCE, etc) or demographic data in order to better identify and focus investigation towards students in need of support. This could be used to automatically generate lists of potentially at-risk students that could be communicated to a subject's examiner at the start of a semester to reduce manual workload.

References

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

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences*, *11*(21). https://doi.org/10.3390/app112110007

Bottcher, A., Thurner, V., & Hafner, T. (2020). Applying Data Analysis to Identify Early Indicators for Potential Risk of Dropout in CS Students. *EDUCON*, 827–836. https://doi.org/10.1109/EDUCON45650.2020.9125378

Dabiran, N., Robinson, B., Sandhu, R., Khalil, M., Poirel, D., & Sarkar, A. (2023). Sparse Bayesian neural networks for regression: Tackling overfitting and computational challenges in uncertainty quantification (arXiv:2310.15614). arXiv. http://arxiv.org/abs/2310.15614

Han, M.-P., Doan, T.-T., Pham, M.-H., & Nguyen, T.-T. (2024). Automation Process for Learning Outcome Predictions. *International Journal of Advanced Computer Science & Applications*, *15*(2). https://doi.org/10.14569/IJACSA.2024.0150291

Hohmann, E., M. D. ,. Ph. D. ,. F. R. C. S., Wetzler, M. J., M. D., & D'Agostino, R. B., Ph. D. (2017). Research Pearls: The Significance of Statistics and Perils of Pooling. Part 2: Predictive Modeling. *Arthroscopy*, *33*(7), 1423–1432. https://doi.org/10.1016/j.arthro.2017.01.054

Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Prog Artif Intell* **5**, 221–232. https://doi.org/10.1007/s13748-016-0094-0

Limanto, S., Buliali, J. L., & Saikhu, A. (2023). Effects of Training Data on Prediction Model for Students' Academic Progress. *International Journal of Advanced Computer Science & Applications*, *14*(7). https://doi.org/10.14569/IJACSA.2023.0140754

Merrick, L., & Taly, A. (2020). The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 17–38). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_2

Rahul, & Katarya, R. (2023). A Systematic Review on Predicting the Performance of Students in Higher Education in Offline Mode Using Machine Learning Techniques. *Wireless Personal Communications*, *133*(3), 1643–1674. https://doi.org/10.1007/s11277-023-10838-x

Saha, A. K., Sharma, A. K., Sahoo, S., Hussain, S. E., & Sahoo, N. K. (2023). Machine Learning Based Prediction of Student's Performance Based on Psychological and Behavioral Data. *Mining Intelligence and Knowledge Exploration* (Vol. 13924, pp. 396–408). Springer. https://doi.org/10.1007/978-3-031-44084-7_37

Senders, J. T., Staples, P. C., Karhade, A. V., Zaki, M. M., Gormley, W. B., Broekman, M. L. D., Smith, T. R., & Arnaout, O. (2018). Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurgery*, *109*, 476-486.e1. https://doi.org/10.1016/j.wneu.2017.09.149

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

Shapley, L.S. (1953). A value for n-person games, *Contributions Theory Games* (Vol. 2).

Zhuoyang, L. & Feng, X. (2023). Interpretable neural networks: principles and applications. *Frontiers in Artificial Intelligence*. 6. https://doi.org/10.3389/frai.2023.974295

Copyright statement