# Predicting school performance using a combination of traditional and non-traditional education data from South Africa

Henry Wandera
University of Pretoria
Email: u17253129@tuks.co.za

Vukosi Marivate
University of Pretoria
Email: vukosi.marivate@cs.up.ac.za

Moinina David Sengeh
Directorate of Science Technology and Innovation
Sierra Leone
Email: dsengeh@statehouse.gov.sl

***Abstract*—The application of big data analytics in education is transforming learning, teaching and administration in schools. Current Education Data Mining (EDM) research focuses on teaching and personalized learning in higher institutions mostly in western countries with limited research conducted in African countries. Most research has been conducted using small datasets, simple learning analytics techniques and machine learning black box models to predict students' performance. Black box modelling approaches use complex structures which are difficult to be easily interpreted by stakeholders. We synthesize EDM approaches and tree based machine learning techniques to identify important features that can predict school performance across African countries such as South Africa. We apply LightGBM a gradient boosting framework and interpretable tree based algorithms on combined data sources from community surveys, school master lists and examination results to perform feature importance. The challenge faced in EDM research is limited education data sources, we merged different existing datasets from government reports and archives. We used community survey data to determine the standards of living in secondary schools within those communities. Cell phone internet, toilets, security, usable water sources, number of teachers and students, school location, and family head were identified as control variables impacting the attainment of schools. LightGBM, underlies the developed prediction model. It empowered the model with high accuracy, stability and easy interpretation hence outperforming XGBoost, decision tree and random forest algorithms.**

***Index Terms*—performance prediction, school performance, national statistics**

## I. Introduction

The deluge of data associated with its value extraction practices can support decision making processes in education. Education big data can be used to address three scenarios, namely, supporting learning, teaching and administration [1]. EDM is an emerging education field that deals with the application of data mining, machine learning and statistics for purposes of understanding students and enhancing their learning environments [2]. This new research paradigm has little research conducted and a limited scope of big data application due to lack of coherent education big datasets, limited set of inclusive big data tools, and use of simple data processing methods caused by limited data science skills [3–5]. Previous studies use learning analytics and traditional statistical modeling methods such as linear regression or logistic regression which have a linear decision surface, work better with only correlated variables and lack an established paradigm for optimizing performance prediction [6]. Poor outcomes would be predicted if researchers fail to identify all the relevant independent variables and data distribution functions. Other statistical methods such as support vector machines, neural networks have the ability to learn and model non-linear and complex relationships or boundaries but are difficult to interpret [2]. End users such as policy makers prefer prediction models that can not only provide actionable insights but are also easy to understand. Moreover, most EDM research focuses on students' learning activities to predict performance using data sources from socio-economic background, classroom attendance and interactions with learning management systems. An interpretable model built on these datasets can easily be validated by school administrators. It also helps them to provide proactive feedback and suggested resources to particular schools or students. We explore different non-traditional data sources from previous examination results, school features such as location, water availability, internet, household goods and municipality difficulties in South Africa. We do not consider individual learning factors such as motivation, intellectual ability or prior knowledge. We built tree based interpretable models that apply 'if-then' rules as they are easy to be interpreted by policy makers. LightGBM, underlies the developed final prediction model as it outperformed traditional tree based models such as XGboost and random forest. We used SHAP (SHapley Additive exPlanations) to explain model outputs and provide relationships among variables to get better understanding of education systems and the underlying factors influencing performance in schools. The rest of this paper is organized as follows: II discusses related research, background information and limitations. Section III presents dataset description, data sources, limitations and assumptions taken. Section V presents preliminary experimental results and analysis. Section VI summarizes this study, presents limitations and future research work.

## II. Literature review

Most EDM research conducted focuses on predicting students' performance, learning abilities and examination scores using data generated from learning activities, educational system interactions and socio-economic backgrounds. In this

section we discuss similar work conducted using machine learning techniques to provide a research background for our study. The authors in [7] successfully conducted linear regression analyses on digital textbook usage data generated by 233 students to predict course final grades. Their study suggests that digital textbook analytics plays an important role in identifying students at a risk of failing the course. Work in [8] used Hadoop and machine learning algorithms, such as neural network, naive bayes, support vector machine to explore data generated by a student management system, educational administration system, and campus card consumption system to predict high risk students in colleges. The study used information such as gender, grade, college entry scores, meals consumption record, use of hot water record and access to various websites. Another set of authors in [9] proposed a study that used situational theory of publics to predict online learning success. While [10] applied web usage mining, decision trees and neural networks algorithms to predict final marks of students that use Moodle courses. These recent research studies demonstrate the strength of machine learning algorithms in ascertaining performance issues in education. However, black box models trained using neural networks are difficult to interpret. Algorithms with linear decision structures have limitations such as assuming linear relationships among variables, being sensitive to outliers and cannot be applied to unevenly distributed datasets with sub-populations. Several researchers have considered 'white-box' interpretable prediction models which can be easily understood, validated and used by policy makers in decision making. Bayesian belief networks in [11], [12] were applied on log data generated by an intelligent tutoring system to predict predict students' success and the likelihood of answering a question correctly. Bayesian modelling techniques facilitate learning about causal relationships between nodes [13] but they are also difficult to interpret because the information content of each variable is represented as one or several probability distributions. This requires teaching conditional probability concepts to the end users of such models before they understand the underlying Bayesian network graphical structures. Unlike other machine learning algorithms, tree based 'if-then' methods mimic the human level of thinking and provide a logical visualization of data which makes it easy to interpret and validate model outputs. Decision trees were used on students internal assessment data to predict their performance in the final exam [14], and to predict students' drop out at Eindhoven University of Technology [15]. Other authors [16] have also applied fuzzy association rule to predict students performance in an E-Learning environment .

The latter research work demonstrate the effectiveness of 'if-then' models giving useful result with accuracies between 60% and 80%. Much as decision trees reduce the ambiguity in model interpretation, they are unstable - a small change in data can lead to a large change in the structure of the optimal tree. We can apply gradient boosting algorithms to train prediction models as an ensemble of weak decision tree models. Extreme gradient boosting (XGBoost) which implements a gradient boosting method uses more accurate approximations to find the best tree model was used in [17] to predict the performance in mathematics based on socioeconomic status in Brazil and in [18] to predict performance and behavioral analysis of student programming ability . The XGBoost model built to predict programming ability proved to be the most efficient, exhibiting an accuracy of 80% for first dataset and 91% for the second dataset. The limitations in their research include failure to interpret the XGBoost model thoroughly and use of gain and weight to measure feature importance in XGBoost. These two importance type parameters produce different feature orderings hence making it difficult to understand which features are exactly more important. In this study, we use a gradient boosting framework called LightGBM released by Microsoft in 2017. LightGBM improves on XGBoost and outperforms it in training speed and the size of the dataset it can handle but their accuracies are comparable [19]. We investigate which of these tree based algorithms performs better. We also use SHAP to show the impact of each feature in predicting school performance. SHAP values are a powerful tool for interpreting tree models and have guaranteed consistency in feature rankings compared to the gain, or split count methods [20].

## III. Data sets

Many research studies uses small datasets and focuses on determining the impact of learning technologies [9], [10], online discussion forums [21], and personalized learning on success [7], [8], [10], [21–23]. These small textual-structured datasets include digital textbook usage data, course data and system data which makes them simple to integrate and explore using traditional technologies. We merged various non traditional data sources with a diverse domain of features, for example, overall school performance dataset, schools master list dataset and 2016 community survey - which provides socio-economic information about different households. Schools located in the same community were assumed to have the same most socio-economic issues. These datasets were statistically significant with less missing values, a large sample size of over 6000 secondary schools and correlations among features. They also contained both qualitative and quantitative data. Qualitative data was good for exploring household open reactions where as quantitative data from schools such as pass rate, number of teachers, was good for answering quantitative questions and can also be used in answering a regression problem.

### *A. Data collection and preparation*

Table I shows the nature of datasets which were collected from four diverse sources. First, the school performance dataset containing final exam results was extracted from the 2018 National senior certificate school performance report published in a pdf file on the department of education site. Second, the locations of schools were found using Google's API. For some of the schools' location that was incorrect (returned null or a location outside of SA), the location of

the district it falls under was used instead. Third, the schools master list datasets for each province were acquired from the department of basic education website in separate files. We downloaded 11 xlsx files and merged them to form a complete master list with 6196 schools across the country. These first three datasets were merged together using EMIS code as the key. Each school as a unique Education Management Information System (EMIS) code. 610 schools with missing attribute values were removed. Lastly, the 2016 community survey dataset was acquired in a csv format from Statistics South Africa (STATS SA) website. This is a large-scale survey that targeted approximately 1.3 million households with the objective of providing population and household statistics at municipal level to government and the private sector to support planning and decision-making. The community survey data was merged with schools data (school performance, locations and master list) using local municipality names as keys. We investigate how these features from community survey affect the performance of schools within local municipalities. Assumption: For community survey dataset, we calculated the most frequent value for every attribute in all distinct municipalities and assigned it to all the schools within the same local municipality. This means that if a municipality had the most number of households with no televisions or experiencing high crime rates, then the schools in that municipality were assumed not to have televisions and also experience high crime rates especially day schools. This assumption was taken based on the current school quintile system. In South Africa, schools are grouped into 5 categories (quintiles) based on the relative wealth of their surrounding communities. Schools in the poorest communities are classified as Quintile 1 and schools serving the wealthiest communities are classified as Quintile 5. Usually schools in quintile 1, 2 and 3 are no fee schools to a greater extent supported by the government.

TABLE I
NATURE OF THE DATASETS

| Datasets (sources) | Attributes | Merging keys |
|---|---|---|
| School performance | EMIS, Name, Province, District, Quintile, 2016, 2017 and 2018 exam % Achieved | EMIS |
| School locations | EMIS, Province, latitude, longitude, Quintile, district longitude and district latitude | EMIS |
| Schools master list | School name, EMIS, sector, specialization, ownerland, ownerbuild, school latitude and longitude, District name, Local municipality name, Rural_Urban, Quintile | EMIS, District name, Local municipality name |
| 2016 Community survey | Household goods such as TV, radio, microwave, cellphone, gas stove, family head (male or female), age of the family head, Local municipality difficulties, Safety in day and night, rate electricity, rate water, rate hospital, District name, Local municipality name | District name, Local municipality name |

The final dataset has 5586 secondary schools and 73 attributes. The exam achieved percentage for each year (2016, 2017 and 2018) means the pass rate. It was calculated from the number of students who passed over the total number of students who sat for the exam in that school. Ownerland means owner of school land, ownerbuild means owner of school building, Rural_Urban records whether the school in a rural or urban area, rate hospital, electricity and water contains ratings by community members for particular institutions, for example, good, average and bad. Other attributes are obvious such as school name, province, district and households goods.

## IV. METHODOLOGY

In this study, we use tree based algorithms as they are proved to be more easily interpreted by end users of the models. These include decision tree, random forest, XGBoost and LightGBM algorithms as shown in Figure IV. We investigate their performances on the same training dataset and testing dataset.

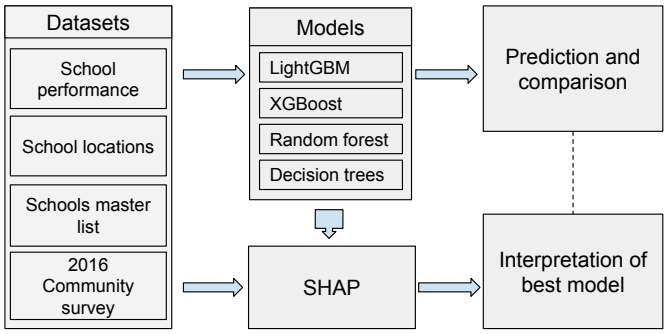


Fig. 1. Proposed approach. SHAP accepts a model as a parameter input for its explainer function to create an object that is used to calculate shap values of the dataset

### A. Data preprocessing

We created classification labels using the average pass rate for every school. We calculated the average of pass rates by ignoring the blank cells in case the school didn't have pass rate results for any of the years. If the average pass rate is greater than or equal to 50%, a school was labelled 1 to indicate a good performance else 0 for poor performance. We created dummy variables for every categorical features. The dataset was split in to training set and testing set on a ratio of 75% to 25% respectively. The training set has 4189 observations while the test set has 1397 observations.

### B. Supervised classification

We apply supervised classification algorithms to train models to learn the mapping function from the input features to the pass rate labels. The testing set was used to test the prediction performance of the models trained by different algorithms. To improve the precision, we used a stratified 10-Folds cross-validator to train and evaluate our models on every strata. We compared four classification algorithms described below:

Decision trees [24], [25]. Decision tree algorithms mimic human level of thinking by representing attributes in a tree-like

structure. Each node represents an attribute, each link between nodes represents a decision rule and the output is represented by the leaf nodes. The best attribute in the dataset is placed at the root of the tree. The training set is then split into subsets. Subsets are made in such a way that each subset contains data with the same value for an attribute. These steps are then repeated for each subset until leaf nodes are found in all the branches of the tree. They are unstable, prone to overfitting and require taking optimal choices at every node.

Random forest [26]. This is a bagging algorithm that uses ensemble learning techniques by training every tree independently and collecting the various decision trees whose results are aggregated into one final result through voting mechanisms. Random forests outperform a single decision tree by solving the problem of overfitting and reducing variance.

XGBoost algorithm [27]. It applies bagging and boosting technique. It trains the different models by resampling the data but subsequent models are trained while taking errors made by previously trained models into account. This is called boosting and it reduces the bias where as bagging reduces variance.

LightGBM [28]. It uses a boosting technique and outperforms XGBoost in training speed and the size of the dataset it can handle but their accuracies are comparable. It is optimized to support parallel learning with higher efficiency while handling big datasets. LightGBM grows trees leaf-wise considering the leaf with max delta loss which reduce more loss than the level-wise algorithms hence better accuracy. However, it may cause to overfitting when a small dataset is used, this is overcome by setting the max_depth to specify the depth of the splits.

### C. Approaches to feature importance

We apply the same machine learning classification algorithms to perform feature selection because they have inbuilt functionality to examine important features. For instance, XGBoost plot importance function takes into account the features gain, cover and weight while evaluating the number of times a feature causes the tree to split. However, when you set the importance type parameter type to either weight or gain, the order (significance) of the features is not stable and consistent [29] which makes it hard to identify the exact significant features.

We use SHAP values method [29] to interpret the results of these tree models because they offer guaranteed consistency in feature rankings. This is because each run returned different top 30 features. So doing it 10 times while taking the mean magnitude of the shap values returned and then sorting at the end meant top 30 features would always be at the top irrespective of the number of splits. SHAP values also can also explain individual predictions by being able to interpret the impact of having a certain value for a given feature in comparison to the prediction we would make if that feature took some baseline value. For example, how much was a prediction driven when the number of teachers was set to 14.

## V. Preliminary results and analysis

Table II shows the final school performance prediction results. One of the objective of this study was to compare the performance of LightGBM with the prediction performance of benchmark models. We used stratified 10-folds cross validator to shuffle our training data and then split it into 10 splits and each split was used as a test set. All the models were trained on the same training dataset and tested on the same test sets. We performed parameter tuning to all classifiers in order to get optimal results. We also trained 10 models for every algorithm and considered the average of the evaluation metrics.

TABLE II
Final school performance prediction results

| | **Lightgbm** | **xgboost** | **random forest** | **decision tree** |
|---|---|---|---|---|
| **Accuracy** | 81.7% | 88.8% | 87.8% | 63.7% |
| **Sensitivity** | 82.4% | 98.7% | 94.1% | 61.9% |
| **Specificity** | 75.5% | 22.5% | 36.4% | 78.1% |
| **AUC** | 79.0% | 87.1% | 65.2% | 70.0% |

### A. LightGBM

According to accuracy and the trade-off between specificity and sensitivity, LightGBM was considered the best algorithm for our dataset with an accuracy of 81.7%, sensitivity of 82.4%, 75.5% and Area Under Curve (AUC) of 79.0%. Models with AUC closer to 100% have good measure of separability. The XGBoost model outperforms all other algorithms in all cases except for specificity. It has specificity of 22.5% which means it is good at predicting only good performing schools but not poor schools. According to this classification problem, a higher specificity value would be preferred in case of identifying schools that are at a risk of failing students. Since we seek to propose interventions such that we can increase the performance of students by providing key performance control features, a model that has the potential negatives cases is preferred. The random forest model has an accuracy of 87.7% and sensitivity of 94.1% which are higher than those for LightGBM and decision tree. However, it also best for predicting good performing schools. If we are to consider models that are capable of predicting all cases while maintaining the level of accuracy, LightGBM emerges best followed by decision tree.

### B. Feature importance

Figure 3 shows the top 30 features that were used most in predicting schools performance. We used the SHAP summary plot built in function to plot the mean absolute value for each feature. SHAP values represent a feature's responsibility for a change in the model output. From Figure 3 we can learn that number of educators contributed most followed by number of learners. This means that teacher-student ratio has an impact on performance. The location of the school either in urban or rural areas plays an important role. The sex of the family head and their age has a great impact in our model performance.

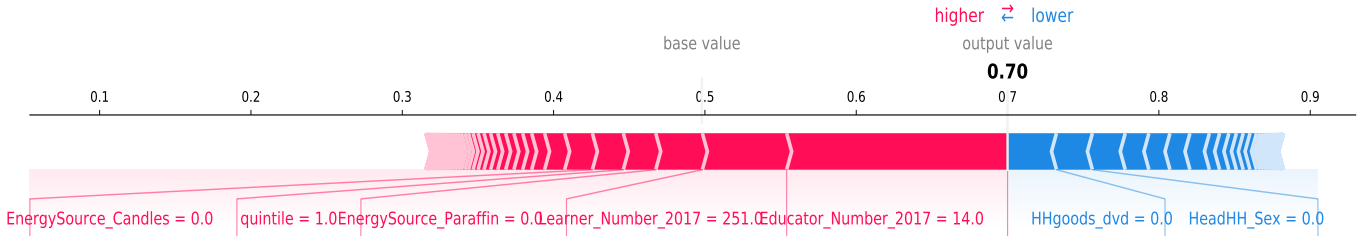


Fig. 2. A force plot showing a single prediction of 0.7 and the features controlling the model output. The red features push for a higher prediction while blue features decrease the prediction

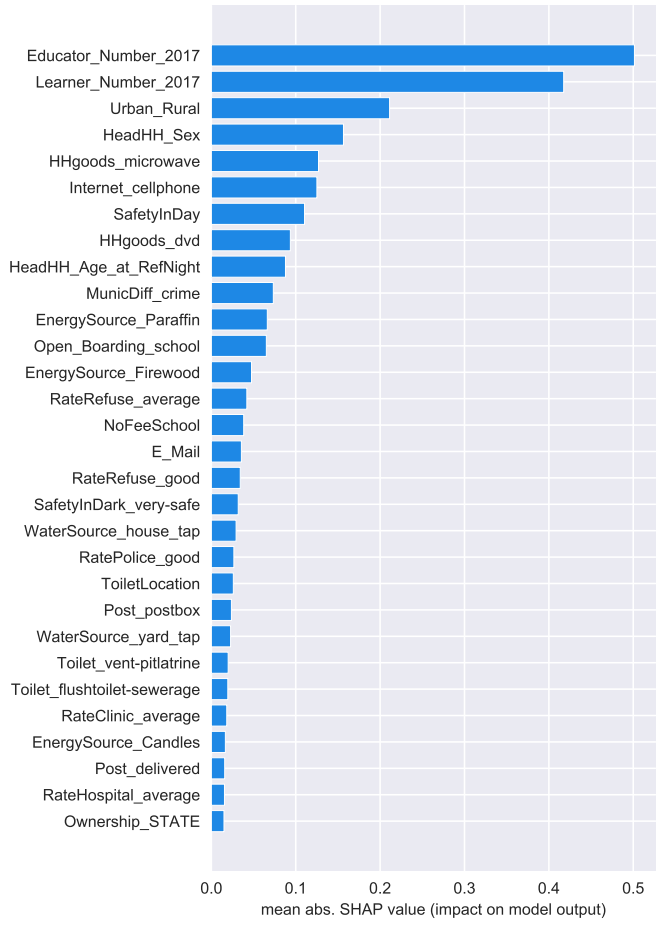


Fig. 3. Feature importance: A bar plot showing top 30 selected features ranked in the order of mean absolute value of the SHAP values (average impact on model output magnitude)

### C. Model interpretation

Figure 2 shows that the school is 70% likely to have its pass rate greater than 50%. For this prediction, the base value is 0.5 and the output value is 0.7 (good performing school). Using SHAP, we are able to interpret and visualize

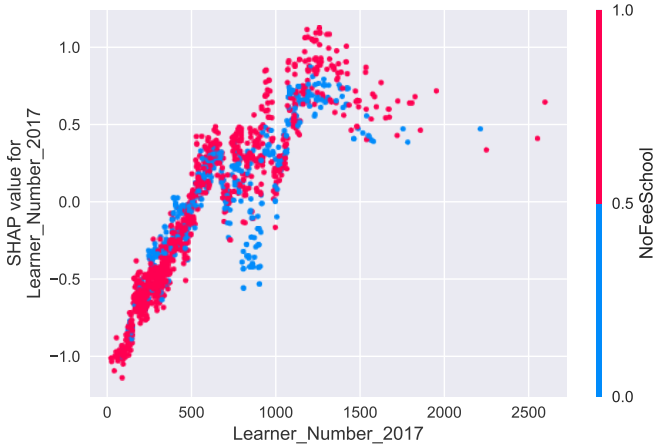


Fig. 4. Dependence plot showing a change in the prediction of pass rate as the number of learners change in both fee and no fee schools. The number of learners (over 500) in no fee schools has a positive impact on performance

the contribution of features. Features on the right tend to push the model prediction to the base value while those on the left push the prediction to the output value. In this case, educator number provides the biggest impact as 14, however if most of the students come from households with no dvds and where the house head is a female, then it has a meaningful effect on reducing the output. Therefore, end users of this model are able to interpret and understand the impact factor of all the control features. To understand the effect of one feature in the prediction, we can plot the SHAP value of that feature against other feature SHAP values in the dataset as shown in Figure 4 From Figure 4, we can learn that a change in the number of learners has a great impact on the performance of the school across schools both fee and no fee schools.

## VI. Conclusion and Future Work

This paper presents our preliminary research and describes the importance of tree based machine learning algorithms and their application predicting school performance. We compared the performance of LightGBM with the prediction perfor-

mance of benchmark models trained using XGBoost, random forest and decision trees. Results from the experiments show that LightGBM model emerged best with it's results being comparable to XGBoost. We used SHAP to interpret our model such that it can be easily understood or validated by end users. The proposed tree based algorithms have limitations. They require correct parameter tuning, LightGBM leaf-wise growth may be overfitting if wrong parameters are used, for example, large number of leaves may cause overfitting and parameters are dependent on each other. We used existing datasets which didn't have most of the school features such as libraries, laboratories and learning materials. In future, we consider using other existing datasets with more features and also expand the scope of study to include other African countries. We are currently exploring schools census data acquired from Sierra Leone and we look forward to using the same machine learning algorithms to extract insights that may be valuable to school administrators. Our focus is to expand the scope of EDM and also provide actionable insights and models to improve education across Africa.

## REFERENCES


[1] B. Daniel, "Big data and analytics in higher education: Opportunities and challenges," *British journal of educational technology*, vol. 46, no. 5, pp. 904–920, 2015.

[2] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[3] S. Buckingham Shum, M. Hawksey, R. S. Baker, N. Jeffery, J. T. Behrens, and R. Pea, "Educational data scientists: a scarce breed," in *Proceedings of the third international conference on learning analytics and knowledge*. ACM, 2013, pp. 278–281.

[4] I. Koprinska, J. Stretton, and K. Yacef, "Students at risk: Detection and remediation." in *EDM*, 2015, pp. 512–515.

[5] B. K. Daniel, "Big data and data science: A critical review of issues for educational research," *British Journal of Educational Technology*, 2017.

[6] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Computers in Human Behavior*, vol. 47, pp. 168–181, 2015.

[7] R. Junco and C. Clem, "Predicting course outcomes with digital textbook usage data," *The Internet and Higher Education*, vol. 27, pp. 54–63, 2015.

[8] Y. Xiaogao and P. Ruiqing, "Research on big data-driven high-risk students prediction," in *Cloud Computing and Big Data Analysis (ICCCBDA), 2017 IEEE 2nd International Conference on*. IEEE, 2017, pp. 145–149.

[9] M. J. Kruger-Ross and R. D. Waters, "Predicting online learning success: Applying the situational theory of publics to the virtual classroom," *Computers & Education*, vol. 61, pp. 176–184, 2013.

[10] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use moodle courses," *Computer Applications in Engineering Education*, vol. 21, no. 1, pp. 135–146, 2013.

[11] Z. A. Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan, "The effect of model granularity on student performance prediction using bayesian networks," in *International Conference on User Modeling*. Springer, 2007, pp. 435–439.

[12] Z. Pardos, N. Heffernan, C. Ruiz, and J. Beck, "The composition effect: Conjuntive or compensatory? an analysis of multi-skill math questions in its," in *Educational Data Mining 2008*, 2008.

[13] D. Heckerman, "A tutorial on learning with bayesian networks. microsoft research," 1995.

[14] S. A. Kumar *et al.*, "Efficiency of decision trees in predicting students academic performance," 2011.

[15] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study." *International Working Group on Educational Data Mining*, 2009.

[16] A. Nebot, F. Castro, A. Vellido, and F. Mugica, "Identification of fuzzy models to predict students performance in an e-learning environment," in *The Fifth IASTED international conference on web-based education, WBE*, 2006, pp. 74–79.

[17] B. Stearns, F. Rangel, F. Rangel, F. F. de Faria, J. Oliveira, and A. A. d. S. Ramos, "Scholar performance prediction using boosted regression trees techniques," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2017.

[18] M. Sagar, A. Gupta, and R. Kaushal, "Performance prediction and behavioral analysis of student programming ability," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2016, pp. 1039–1045.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[21] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers & Education*, vol. 68, pp. 458–472, 2013.

[22] S. Uddin, K. Thompson, B. Schwendimann, and M. Piraveenan, "The impact of study load on the dynamics of longitudinal email communications among students," *Computers & Education*, vol. 72, pp. 209–219, 2014.

[23] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining social media data for understanding students learning experiences," *IEEE Transactions on Learning Technologies*, vol. 7, no. 3, pp. 246–259, 2014.

[24] J. R. Quinlan, "Simplifying decision trees," *International journal of man-machine studies*, vol. 27, no. 3, pp. 221–234, 1987.

[25] ——, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[28] H. Shi, "Best-first decision tree learning," Ph.D. dissertation, The University of Waikato, 2007.

[29] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.