Thuto: Depth Analysis of South African and Sierra Leone School Outcomes using Machine Learning

by

Henry Wandera

Submitted in partial fulfillment of the requirements for the degree
Master In Information Technology (Big Data Science)
in the Faculty of Engineering, Built Environment and Information Technology
University of Pretoria, Pretoria

December 2020

Publication data:

Henry Wandera. Thuto: Depth Analysis of South African and Sierra Leone School Outcomes using Machine Learning. Masters Dissertation, University of Pretoria, Department of Computer Science, Pretoria, South Africa, December 2020.

Electronic, hyperlinked versions of this thesis are available online, as Adobe PDF files, at:

https://dsfsi.github.io/

https://orcid.org/0000-0001-6270-217X

Thuto: Depth Analysis of South African and Sierra Leone School Outcomes using Machine Learning

by

Henry Wandera
E-mail: jajawandera@gmail.com

Abstract

Available or adequate information to inform decision making for resource allocation in support of school improvement is a critical issue globally. In this paper, we apply machine learning and education data mining techniques on education big data to identify determinants of high schools' performance in two African countries: South Africa and Sierra Leone. The research objective is to build predictors for school performance and extract the importance of different community-level and school-level features. We deploy interpretable metrics from machine learning approaches such as SHAP values on tree models and Logistic Regression odds ratios to extract interactions of factors that can support policy decision making. Determinants of performance vary in these two countries, hence different policy implications and resource allocation recommendations.

Keywords: Education, Policy-making and Machine learning.

Supervisors : Dr. Vukosi Marivate
Department : Department of Computer Science
Degree : Master of Information Technology (Big Data Science)

Acknowledgements

I would like to express my appreciation to the following people and organisations for their continued support throughout this degree:

- MasterCard Foundation Scholars Program, for all the financial support. They believed in my cause as a vibrant youth and unlocked the potential of a good leader in me. I can only give back by establishing and engaging in activities that will change the communities we live in.
- My supervisor, Dr. Vukosi Marivate, and Dr David Moinina Sengeh for the guidance, timely and appropriate feedback that enabled me progress in this degree. I am grateful for their empathy and connections to new opportunities that enhanced my skill set.
- The Sydney Brenner Institute for Molecular Bioscience family, for their cheerful hearts and allowing me to exercise my skills with them. God led me to this family in the most difficult time where I surely needed both financial and emotional support.
- My family, for helping me to fight against all the odds in my life. I am proud to have a protective and lovely family.
- My friends, for their advise, feedback and memories spent with me. They played with me. They inspired me. They were always showing me the right path.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Businesses are dependent on analytics to mine insights about prevailing situations, and to discover useful trends in the market for purposes of maximising profits, optimising operations and creating new opportunities for their customers. This process has been supported by the effective utilization of streams of data that originate from their customer base. Likewise, Schools, Universities and Colleges generate huge amounts of data (big data) about their environment, students, teachers and their technology usage. Big data means large and disparate volumes of data generated by people, applications and machines in various fields including education. Big data in education can potentially improve learning, help to track progress, success and achievements in educational institutions. It would be insightful to connect various educational data points, for example, data about students' behaviour, examination results, teachers data, school environment data and surrounding community data to improve resource allocation and operational effectiveness in schools.*

In education, analysing schools data can support policy formulation with the goal of addressing three scenarios, namely, supporting learning, teaching and administration [16]. This research discusses features that administrators should consider providing and maintaining in schools. We also use these significant features to predict whether if made

available, a school will perform well or not.

We apply interpretable machine learning techniques on existing education data to extract essential factors that can improve the predictions of school learning outcomes. Through this machine learning approach, we are able to illuminate the likely factors that policy-makers and school stakeholders should consider when seeking to improve their education outcomes. We make suggestions for education policy after comparing and distinguishing our results to factors that have been identified from research applied in other countries using methods such as qualitative surveys.

We use a diverse range of existing datasets including the national examination outcomes, basic features of schools and communities, and national statistics data, specifically for schools in South Africa (SA) and Sierra Leone (SL). South Africa is a country on the southernmost tip of the African continent while Sierra Leone is a country in West Africa, on the Atlantic Ocean. These countries differ geographically, economically and socio-politically.

We chose these because by the time of this study, they had existing education datasets, and initiatives of classifying their schools to improve educational resource provision and governance. In 2016, South Africa department of education introduced a school classification system based on performance and not social class to improve the allocation of resources in schools [46]. In 2018, the government of Sierra Leone embarked on a new program that employs data science methods to explore survey data collected from schools to facilitate planning and budgeting, and providing proper service delivery [50].

1.1 Motivation

In some schools there is not enough information to act as a guiding resource when developing educational infrastructure and allocating school supplies that can be potentially improve performance.

Unlike the educational conditions in some developed countries, education institutions in African developing countries continue to face underlying challenges with access and quality such as high rates of drop out, lack of or poor quality of learning and teaching resources, lack of enough qualified and experienced teachers, and poor basic level infrastructure required for every school to function properly and perform well.

This work investigate and compare the role of school and community level features in two African education systems in Southern Africa and Western Africa. The results can be used by school administrators to improve resource allocation in schools. Education policymakers can also use these results to formulate legally-binding standards every school should meet for suitable learning to take place.

Similar work in [4] conducted by various researchers with different research objectives and educational datasets in the field of Educational Data Mining (EDM). EDM is a field that applies statistical, machine-learning, and data-mining algorithms over the different types of educational data [9]. It aims to analyze education data to resolve educational research issues. EDM research work has been applied in the management of classes, monitoring the effectiveness or usage of educational technologies and content, and predicting students' performance [62]. Previous EDM work predominantly used small datasets from sources such as data from e-learning platforms, learning management and intelligent tutoring systems and it is mostly conducted in developed western countries. Firstly, we use variables from school-level and community-level data sources to identify how and to what extent school performance is related or affected by these variables. Secondly, this is an African based study focusing on South Africa and Sierra Leone education systems. For interpretability, we use Logistic Regression odds ratios and tree-based machine learning models that mimic the human level of decision making to decode model results and determine the likelihood of a factor being critical in influencing the school performance.

1.2 Objectives

The main objective of this study is to investigate and identify determinants of school performance from both school and community level features. These factors can help to determine the optimal location of new schools, the optimal allocation of teachers and provide data-driven answers to direct where to provide specific infrastructure such as electricity, water or latrines. The following are the questions this study seek to address:

1. Does the school environment influence performance of students in schools?
2. Does the community environment influence performance of students in schools?
3. Which school features are most relevant in determining overall school performance?
4. What are the characteristics of the best performing schools in each country?
5. How do school features in Sierra Leone and South Africa differ in determining performance?

1.3 Contributions

This research study contributes the following:

1. The collection of existing education datasets in South Africa and merging them with the 2016 community survey data.
2. An analysis of merged South Africa datasets to discover the impact of community-level features and household goods on students' performance in schools.
3. The extraction of insights from the Sierra Leone 2018 school survey data to investigate the association of school-level features and school performance.
4. Prediction of school performance and categorizing of schools for affirmative action especially those that require an increase in opportunities and resources.

5. A policy implication discussion highlighting key points which school stakeholders should consider during policy formulation and resource allocation, for example, is it more important to have playgrounds or to have running water in schools?

6. Provide an information base to act as a guiding tool in directing the optimal location of schools and what basic infrastructure to provide.

1.4 Derived Publications

The following publications were written from the research work of this study:

- Wandera Henry, Vukosi Marivate, and David Sengeh. "Investigating similarities and differences between South African and Sierra Leonean school outcomes using Machine Learning." *arXiv preprint arXiv:2004.11369 (2020)*, pages 131-–136, 2019.

- Wandera Henry, Vukosi Marivate, and Moinina David Sengeh. "Predicting National School Performance for Policy Making in South Africa." *In 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pp. 23-28. IEEE, 2019.

1.5 Dissertation Outline

The rest of this study is organized as follows:

- **Chapter 2** consists of a comprehensive review of other related conducted studies. This chapter situates this study within an existing body of knowledge and provides details of the theoretical framework and methodology developed.

- **Chapter 3** explains the technical details of the methodological approach and algorithms used in this study.

- **Chapter 4** discusses the various types of datasets used in this study. This chapter discusses the data collection and preprocessing steps taken.
- **Chapter 5** summarizes the main characteristics of the datasets used to provide baseline insights into what the datasets could tell beyond the formal modelling process.
- **Chapter 6** presents the predictive results from the experiments and their interpretation.
- **Chapter 7** presents recommendations and the policy implications of the identified factors.
- **Chapter 8** summarises the findings and discusses the limitations of this study. A brief discussion is also presented on future work that can extend on the work in this study.

Chapter 2

Literature Review

Schools are established to impact knowledge in their students, whose combined learning results represents the performance of the school. There are many factors such as school management, experienced and qualified teachers, and other school resources that play an influential role in determining the performance of their students. Understanding which performance determinants are critical in a school is a great deal of concern especially to governments and school management teams. Well-performing schools understand their strengths and weaknesses and continue to capitalize on this awareness to maintain their performance or gradually improve their performance with time. In most low-performing schools, there is still lack of enough information about which fundamental variables can potentially change their poor results. This study investigates determinants of performance using existing data in South Africa and Sierra Leone schools to inform policy making and school stakeholders.

This chapter discusses students performance related studies that were globally conducted to find the influence of various factors and environments on learning outcome. Section 2 introduces the need for this type of research and how they differ in context. Section 2.1 presents various research work conducted to predict performance at student level using data generated by personalized digital learning platforms. Section 2.2 discusses prior work conducted focusing on the role played by learning environments in determining

the academic performance of students. These learning environment factors consists of school facilities, security and safety, parents and teachers interactions. This study mainly focused on predicting school performance using interpretable machine learning techniques by looking at how various learning environments influence performance, therefore in Section 2.3 different education data mining approaches used in predicting performance are briefly discussed. Lastly, a summary of this chapter is presented in Section 2.4.

Measuring schools' performance and investigating the determinants of this performance is paramount. Schools without performance measurement systems are similar to organisations that operate without purpose and direction. Schools that exist have to disseminate knowledge and impact learning conduct assessments to evaluate the performance their students. The purpose of measuring performance is not only to know how students are performing but also to enable it to perform better. In order to improve performance, there is need to first understand the role played by various factors in and around the school. These factors will further induce the creation of performance matrices to measure, monitor and evaluate school achievements.

School performance is a factor of multiple variables, such as availability and effective use of school resources, student leadership, teachers, school management, role of parents/guardians, non-academic staff, security, surrounding community and the government. Effectively, if the school can have strategic purpose, create a clear definition of performance, and demarcate the roles of every stakeholder, they can achieve their defined performance objectives.

Many recent studies have focused on the addressing the problem of performance in schools focusing on predicting students' examination scores, understanding learning abilities using data generated from on-site school interactions, digital learning activities, school environment and socio-economic backgrounds [63]. Subsequent subsections discuss various related research that used different approaches to address performance issues both at student level and school level.

2.1 Student-level performance

Digital technology has drastically changed almost every dimension of human life ranging from various communications types to banking, entertainment and shopping. Likewise, in education, digital tools have been deployed to personalise how students learn and how teachers teach through the use of virtual electronic learning systems or digital learning platforms. This is a type of learning supported by technologies to deliver instructional materials and content to learners. It can be a combination of blended learning, virtual reality, electronic textbooks, electronic learning and digital environments tailored to facilitate learning. In this section, various prior research is discussed to ascertain the impact of electronic tools and materials on learning success at a more personalized level.

Junco et al in [34] 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. Textbook analytics is a sub-category of learning analytics which is "measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [71]. Learning analytics is an application of big data and predictive analytics in educational settings [22]. Much as digital textbooks analytics in [34] served as standalone predictors of student success, [58] discovered that most students preferred printed out materials when studying because of difficulty reading from a screen. There are many other online materials and systems such as learning management systems schools use to render learning content to their students. Understanding how students interact with them can also provide some insights to help address learning and performance problems. Fore instance, scholars in [84] explored data generated by students on a learning management system. It was found out that quizzes were the most relevant activity as they did not require students to dedicate a lot of time to complete less number of tasks. The results implies that assessment is also one of the main factors of students success.

Assessment is always a crucial process in education, it is a learning tool which helps to measure performance, monitor progress and reveal strong or weak areas of success.

Assessment materials can be used both online or offline. Offline assessment materials are mostly in form of paper-based tests while online assessment materials designed on digital learning platforms. These digital learning platforms make assessment easier to design and manage, and students usually find them easier to engage with. They also have a quicker appropriate feedback time since most of the moderation and marking process can be automated. Digital learning platforms are time saving with the ability to offer the functionality of reviewing correct and incorrect answers, and recording students marks in easily visualized or exported formats. These benefits have attracted more students to have interest in online courses which has also increased the willingness of more schools to offer them. Recent studies have focused on investigating these online courses by predicting the learning success of students that use these digital learning platforms. Scholars in [38] conducted a study on 182 students that completed an online public relations course and found out that with minimised distractions, there was greater student involvement in the class and greater exam scores. Another study in [59] also applied web usage mining techniques to predict final marks of University students that used Moodle courses. Moodle is a free and open-source online learning management system [20]. Results also showed that students who did not pass any quizzes failed in the final exam. Students who did more than 10 assignments and read more messages on the forum obtained excellent marks in their final exam.

Previous studies presented above and in [14, 18, 41, 48, 60, 79] focus on personalized learning by determining the impact of learning technologies, digital textbooks and online discussion forums on success and dropout. Apart from online learning environments, learning also occurs in physical structures such as classrooms, libraries and laboratories. There are many other school environment features and factors which are influential in determining students success. It is important to extensively conduct educational research to cover most aspects under which teaching and learning occurs. In a big data generation, massive amount data are captured from different sources such as online platforms, learning materials and activities, examination results, teachers data, school resources and data related to administrative processes. Collecting and analysing all these data items not only increase the size of available educational data but also yields more useful insights and education research outputs.

Data used in previous research is limited in size and scope because it is collected from a particular learning platform and specific category of students. There is need for additional educational big data research that uses a variety of data sources to solve and address prevailing education problems. This study uses a big data approach by collecting and mining insights from multiple data sources. It also focuses on improving organisational effectiveness which falls under the field of academic analytics as categorized by The Society for Learning Analytics Research [33]. Academic analytics is similar to business analytics and focuses on the improving resource allocation, work flows and processes within education institutions. Strategic resource allocation and improvement of workflows helps to create a positive school environment which encourages students' attendance, reduce stress among students and teachers hence improving teaching and learning. The resources can be within the school or surrounding communities. In order to extend the scope of data sources used in this research, we look at both community household data and school surveys. In the next section, related research is discussed to describe how this proposed research is related to prior research conducted to investigate performance determinants at school, household or community level. We investigate how school-level features, community-level features and digital tools in such as televisions, computers and televisions in households are associated to the overall performance of schools operating within these areas.

2.2 School-level performance

Research shows that positive learning environments can significantly lift students' achievements in schools. Good classrooms, feeding and accommodation are linked to high satisfaction with the schools. Students studying from well stablished safe and clean environments feel protected and ready to learn. Protection can also be inform of health and sanitation where students have access to clean water, latrines, drainage and waste disposal facilities. Since students spend more time at both at school and home, these facilities should be designed and optimised to be conducive and supportive for effective learning to take place since they are crucial in shaping learning experiences and are

linked to high academic performance in [2, 8, 21, 37, 54].

It is also indicated in [2] that school facilities such as laboratories affect academic performance. Students with enough physics laboratory facilities performed better than students in schools with less or without facilities. Availability of well stocked classrooms, libraries and laboratories with adequate facilities gives students an opportunity to learn better and improve their academic achievement [80].

The geographical location and safety in and around schools also has a significant influence on students' performance. According to the US National Centre for Education Statistics report covering topics of school safety such as victimization, use of drugs and bullying, results from 2007 to 2010 data showed that in 2009-10, about 74 percent of public schools recorded one or more violent incidents of crime and 7% of students reported avoiding one or more places in school for fear of their safety [57]. The safety of students at school, home or surrounding community where they meet strangers while traveling to and from school is necessary in encouraging learning. Research in [73] revealed a strong association between measures of school safety and average student achievement in Chicago schools. Students were unable to concentrate on academics when they feared for their physical well-being. These findings emphasize the need to provide safe and secure learning environments. Some schools have guidelines on safety and security for preventing violence, abuse and unauthorized access to school premises. However, research in [5, 15] has found that some students tend to have disruptive behaviour in schools and encourages enforcing school rules. Parents and teachers have been challenged to be at the core of enforcing these rules. Results in [78] indicated that authoritative parenting continues to influence academic performance while in [66], the academic performance of students was linked to the values and aspirations they share with their parents, home learning activities and expectations and interactions they have with their teachers and classmates.

The research work in this study also investigates various determinants of academic performance ranging from safety and security issues, school geographic location, school fence, to school and household facilities such as latrines, access to water, laboratories, ans parents and teachers characteristics. Unlike previous work, a machine learning methodology

is used to extract patterns within the data. The next section discussed how this study applies different education data mining and machine learning methods.

2.3 Education Data Mining and Machine Learning Techniques

Educational data mining refers to the extraction of insights from data generated by learning activities or systems using techniques such as statistics and visualization, clustering, classification, pattern mining and text mining [61]. These techniques can be applied using machine learning algorithms such as decision trees, neural networks and statistical algorithms. Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [44].

In this section, we discuss 2 education research methods: statistical analysis of factors and qualitative methods which have been predominantly used in most educational research.

Qualitative work in [3, 10, 14, 60, 76, 79] monitored students participation in forums, depth interviews and group discussions. This type of work falls in the social context since it studies the subjectivity in students actions, their thinking, and how they react to different educational events.

Qualitative methods help to identify causes and avail us with the information about how various forms of learning occur, and provide solutions to improve students learning. One of the limitations of these methods is that there is they do not provide a mechanism to assist analyse qualitative data mathematically. It is hard to investigate causality, quantify solutions, and verify students' opinions as they have full control over what they provide within those contexts.

The deluge of education data together with increased interest to extract actionable infor-

mation from it has facilitated the use of various statistical analysis methods and improved technologies. These methods sometimes require qualitative methods to investigate some findings further.

Research in [18, 34, 48, 59, 84, 85] used statistical approaches to discover insights and patterns in data from digital learning systems and materials. Some statistical methods for instance that use analysis of variance are prone to outliers. Outliers are also data points but they are usually far from other data points. They can be problematic especially when they cause tests to either miss significant findings or distort real results [70].

Linear regression used in [23, 34] is a statistical analysis method that assesses whether one or more predictor variables explain the dependent variable for instance, is age a predictor of salary. Studies that use Linear Regression assume linear relationships within datasets and all variables to be multivariate normal. In real world, this is not usually the case in many datasets. The correlation is never perfect hence it better to always check the distribution of the dataset using for example, descriptive statistics measures, histograms and scatter plots otherwise these assumption will yield wrong results. Linear regression is also limited to predicting continuous dependent variables which limits its application in solving classification problems.

Neural Networks applied in [59] are suitable for very large datasets but are "black-box" in nature.This makes them difficult to be interpreted especially by the end users or policy-makers. Their study also deemed neural networks to be less useful due to its lack of comprehensibility. The recommended fuzzy rule algorithms applied in [48] and decision trees also applied in [18, 27] because they provided more accurate and comprehensible results which were easy to interpret. By comparison, algorithms like decision trees can easily be interpreted and offer more useful results. Interpretation of machine learning results is critical, for example, a bank won't apply neural networks to predict the creditworthy of a client because they will be compelled to explain to their clients why they did not qualify to get a loan.

This research work focused on interpretability by building models that are easy to use,

interpret and also be validated by policy-makers. In order to deduce more robust research outputs, like we used tree-based models which applies an if-then analysis by mimicking a human approach to decision making and also assumes no existence linear relationships among variables. To avoid limitations of decision tree such as training unstable classifiers, we use scalable and highly boosting decision tree algorithms in [13, 35] and compare the performance with an ordinary decision tree algorithm and a linear classifier - logistic regression which uses a sigmoid function to fit the data.

For interpretability, a Predictive, Descriptive, Relevant framework in [47], SHAP values in [42] and odds ratios in [49] were used to guide the modelling and extraction of patterns from trained performance prediction models. These techniques were also used to visualise significant factors influencing school outcomes. It is easy for end-users to understand results presented in diagrams.

2.4 Summary

Chapter 2 discusses various explicit education data mining research with distinctive research objectives. Some considered predicting learners success in personalized digital learning platforms while other studies focused on the understanding the effect of school facilities, safety and security, and the role of parents and teachers in determining students' performance. These two research paradigms were defined as learning analytics and academic analytics. They used both qualitative and quantitative data mining methods.

For the purpose of this research, an academic analytics approach designed to inform policy making, improve resource allocation and predict school performance using machine learning classifiers was adopted. This study uses quantitative methods like Logistic Regression and tree-based machine learning algorithms. The main advantage of these algorithms lies in their comprehensibility and interpretability. Their results are more useful to end users since they use an if-then analysis. Data from school surveys and community surveys is used. School surveys contained information about the school environment, school facilities, teachers while the community surveys contained data about

household goods.

Chapter 3

Methodology

This chapter provides background information to help the reader understand and evaluate the reliability and validity of the methods chosen to be used in this study. It explains how data was analyzed and processed following a standard process for Data Mining to encourage inter-operable tools across entire data mining process. A discussion of model training algorithms and evaluation techniques used in this study is also presented. Tree based machine learning algorithms, Logistic Regression and SHAP values were used in this research work. Tree based algorithms such as decision trees are always unstable since a small change in data can cause a large change in the structure of the tree [31], but they are more intuitive and easy to explain to non technical stakeholders. SHAP values were used to visualize the modelling results of the algorithms. It should be noted that all the tools and algorithms used were not modified.

The rest of this chapter is organised as follows:

- Section 3.1 presents the implementation details of data mining approach adapted in this study.
- Section 3.2 explains the business understanding phase of this study.
- Section 3.3 presents the data preprocessing tasks that were performed in this study.

- Section 3.4 discusses how the modeling phase and how the algorithms used in this study work.
- Section 3.5 explains how the models were evaluated.
- Section 3.6 presents the deployment phase of CRISP-DM model but focusing on the interpretability of the models that were trained in this study.
- Section 3.7 summarises the contents of this chapter.

3.1 Implementation Details: CRISP-DM model

This study adapted the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model in Figure 3.3. The CRISP-DM model is an open standard process that provides an overview of the life cycle of a data mining project and explains the most used approaches [83]. It comprises of 4 levels of abstraction: phases, generic tasks, specialized tasks, and process instances. There are 6 phases (first level of abstraction) as shown in Figure 3.3, but each phase has list of tasks which are conducted. These tasks are categorised as either generic tasks (second level) or specialized tasks (third level) as shown in Figure 3.2. Generic tasks include series of all possible data mining actions, while specialized tasks describe how actions in the generic tasks should be carried out. The process instance is the last level of abstraction which explains the are criteria, assumptions and techniques which are followed while executing the tasks.

The CRISP-DM model was adopted for its 6 phases (detailed in Figure 3.2) which helped in describing all data mining situations and actions in this study. It provides a high level of flexibility that helps to improve hypotheses and data analysis methods since every phase contributes and double checks the tasks executed in prior phases. The following subsections describe how the 6 phases of CRISP-DM model were applied in this study.

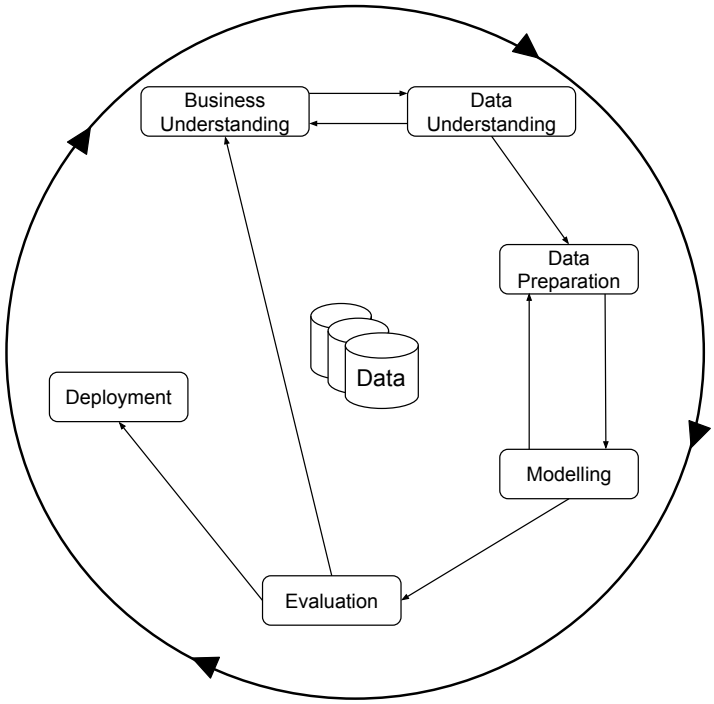


Figure 3.1: Phases of the Current CRISP-DM Process Model for Data Mining

3.2 Business Understanding

CRISP-DM model specifies determining business objectives, assessing the situation and determining data mining goals as generic tasks of this phase. In this study, business understanding referred to comprehension of the problems that were identified by the relevant stakeholders as primary indicators or areas which required investigation in order to improve schools. This included identifying business objectives and the main objective of this study was to investigate determinants of school performance. This phase also included identifying tools and technologies that were used.

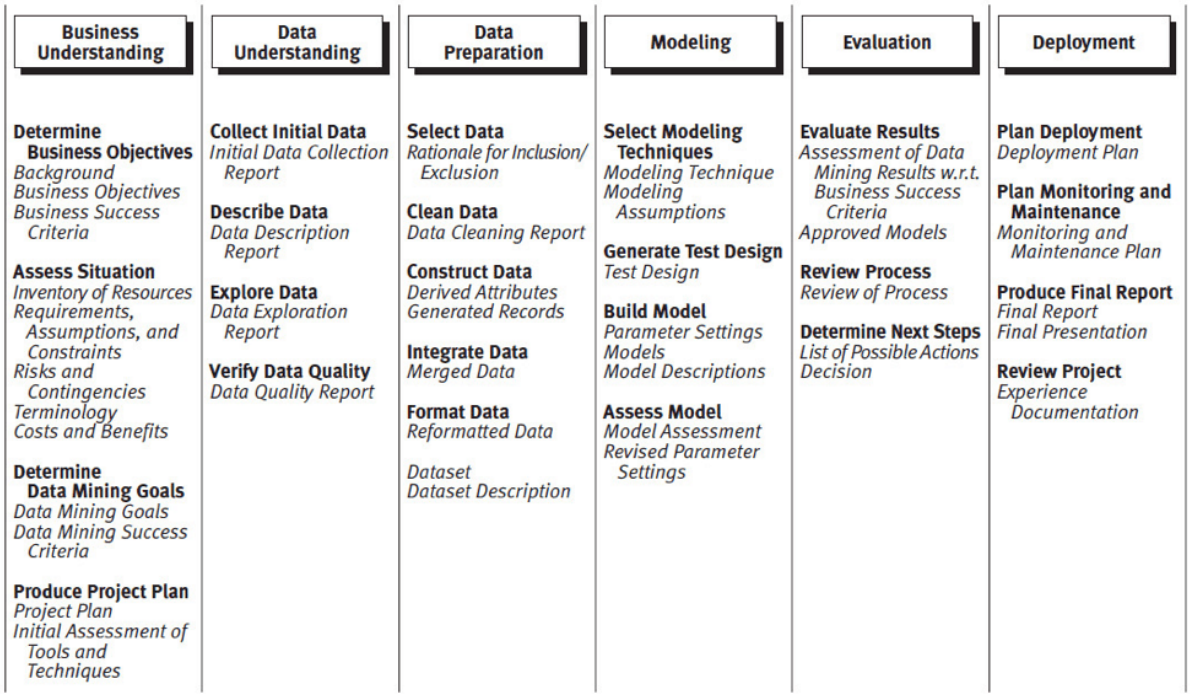


Figure 3.2: Tasks and Outputs of the CRISP-DM Reference Model

3.3 Data understanding and preparation

CRISP-DM splits this item into 2 phases. Data understanding involves data collection, data description, exploration and verification as general tasks. Data preparation involves data cleaning, deriving attributes, formatting and merging data as its general tasks. Chapter 4 and 5 of this study are generic tasks of this CRISP-DM phase. They explain the how data was acquired and explored to extract insights respectively. The following data preprocessing tasks were performed to review and improve data quality, prepare data for exploratory analysis and modelling stage.

- Searching for and correcting inconsistencies in the data.
- Searching for and removing records with spurious data.
- Searching for and removing duplicate records.

- Searching for and redefining or removing outliers.
- Imputing and removing missing values.
- Redefining variables as necessary.
- Calculating new variables like the pass rate (targets for classification).
- Selecting of features that were used in training the models. Multicollinearity cases were identified where some features were highly correlated to each other. Non correlated predictor variables were selected. Each dataset was split into a training set and testing set using a percentage ratio of 70:30 respectively.

3.4 Modeling

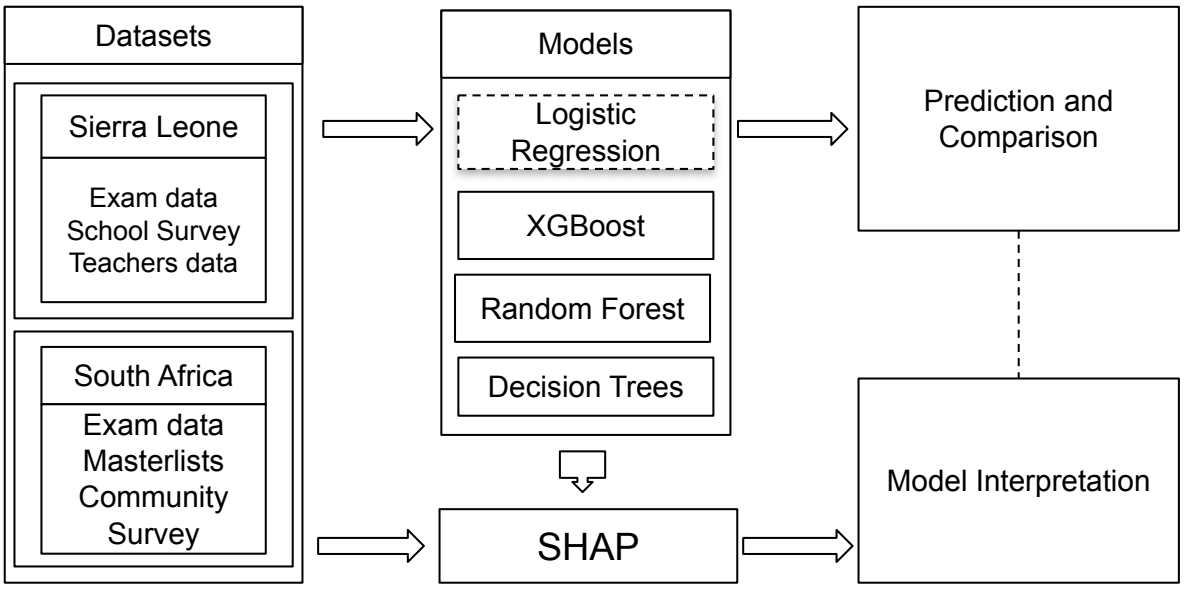


Figure 3.3: Modeling approach

The Modeling phase of CRISP-DM model consists of selecting and applying various modeling techniques, building models, parameter tuning, and assessing model results. In this study, supervised classification algorithms were applied to train different constituent

models for every country using their respective datasets to learn the mapping function connecting input features to the school outcome (pass rate labels). The school outcome for Sierra Leone was measured by the percentage of papers students passed in 2018 WASSCE exams, while for South Africa, it was measured by the percentage of students who passed the 2016 Matric exams.

3.4.1 Classification Algorithms

Classification is the process of grouping ideas and objects into sub-populations [36]. This helps to contrast preset categories for purposes of recognizing their differences. Classification tasks in machine learning refer to predictive modeling problems which receive training data with predefined categories to learn how to predict the likelihood that an instance of data will belong to a particular category. This type of machine learning is called supervised learning where algorithms learn how to map various features to predetermined classes or categories [72]. For instance, classifying if an email is spam or not given some email features. This is an example of a binary prediction task where the classification problem has 2 class labels which are spam or not. Multi-class classification is another type of classification task that has more than two class labels, for example, classifying individual faces and plant species. There are many species of plants, and faces of people hence this type of task requires multi-class classification methods [72, 7].

Binary classification tasks are performed in this study to predict whether the school will fail most students or have most of the students pass their final examination given that school features are in place. It was not possible to perform multi-classification due to limited number of examples in each class, which in turn causes an imbalance classification problem.

The imbalanced classification problem is where the number of examples in each class is unequally distributed. For instance, if the dataset contains 80% of the schools that performed well and only 20% of the schools performed poorly, there will be more cases of well performing schools compared to poor performing schools which affects the per-

formance on some algorithms since tend to predict the majority class in many cases [32]. This study applies the algorithmic approach to solve the class imbalance problem. This problem can be solved by preprocessing methods such as undersampling and oversampling. Undersampling reduces instances of the majority class to balance with those of the minority class, but it has a con of cutting out some relevant dataset features. Oversampling increases instances of the minority class to equate them with the number of instances in the majority class which may lead to overfitting. Overfitting occurs when the model fails to generalize well from the training data to unseen data that is it performs well with high accuracy on the training set but poor when introduced to a new dataset. To address these classification problems, we use ensemble classification algorithms in 3.4.1 to combine predictions of different classifiers and cross-validation that prevents against overfitting by splitting data into K folds and iteratively train each algorithm on k-1 folds while using the remaining fold as the test set.

In order to develop the binary classifiers, two categories or pass rate labels were used in both countries: "fail" and "pass". For SA, schools with pass rate less than 50% were labelled as "fail" and those with pass rate greater or equal to 50% were labelled as "pass". For SL, schools with percentage of papers passed less than 50% were labelled as "fail" and those with percentage of papers greater or equal to 50% were labelled as "pass". The explanatory variables (x) for the models included community level and school level factors such as security, housing, electricity and latrines (see Table 4.1 and 4.2). The dependent variable (y) was the school outcome (pass or fail).

3.4.2 Logistic Regression

Regression analysis tools learn associations within data by mapping input data features with their corresponding continuous numeric output values [67]. Unlike linear regression which predicts theoretically inadmissible values greater than one and less than zero, the Logistic regression predicts probabilities or values in the range of 0 to 1 hence can be used to predict categorical variables [67].

Since this study was not predicting pass rates or the percentage of papers passed but instead identify the likelihood of a school passing or failing, Logistic regression was the most appropriate regression method to apply. It is also a good probability/risk estimator since it provides probabilities and odd ratios to measure the likelihood of an event occurring. This study chose different model parameters maximise the likelihood computed (loss function). A loss function is a measure of fit between a mathematical model of data and the actual data [64]. Parameters are chosen to maximize the goodness-of-fit of the model to the data. A similar study in [11] applied this algorithm and it was able to demonstrate the effects of poor oral health and general health on school performance.

3.4.3 Tree Based Algorithms

Tree based supervised machine learning algorithms were considered in this study because of their high accuracy, stability and ease of interpretation [40]. These included decision trees, random forest and gradient boosting algorithms. Tree based algorithms make splitting decisions by considering the most significant variable to used when splitting the population into two or more homogeneous sub populations. Splitting is the process of dividing a node into two or more sub-nodes [40]. The root node represent the entire population, while the decision node is a sub-node which is used to split into further sub-nodes. The final output is represented by the leaf node. Figure 3.4 shows how tree based algorithms work by depicting the splitting process of deciding whether the school will pass or fail based on its location and number of teachers.

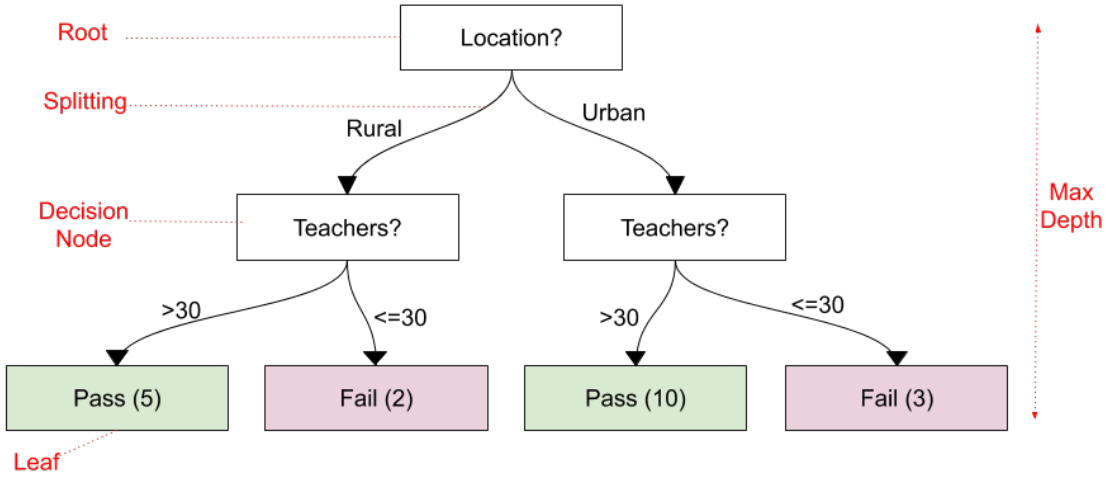


Figure 3.4: A tree-like structure of decision tree algorithms

Decision trees [56, 55] repeats the process of splitting each sub-population until leaf nodes are found in all the branches of the tree. Trees can be are unstable, prone to overfitting if optimal splitting choices are not taken at every node. Decision trees are unstable because they are sensitive to a specific dataset and often change in structure and predictions when presented to a new dataset.

Unlike Decision tree algorithms, Random forest [12] is an ensemble machine learning algorithm which improves performance by combining predictions from multiple independent decision tree models as shown in Figure 3.5. Random forests outperform a single decision tree by solving the problem of overfitting and reducing variance.

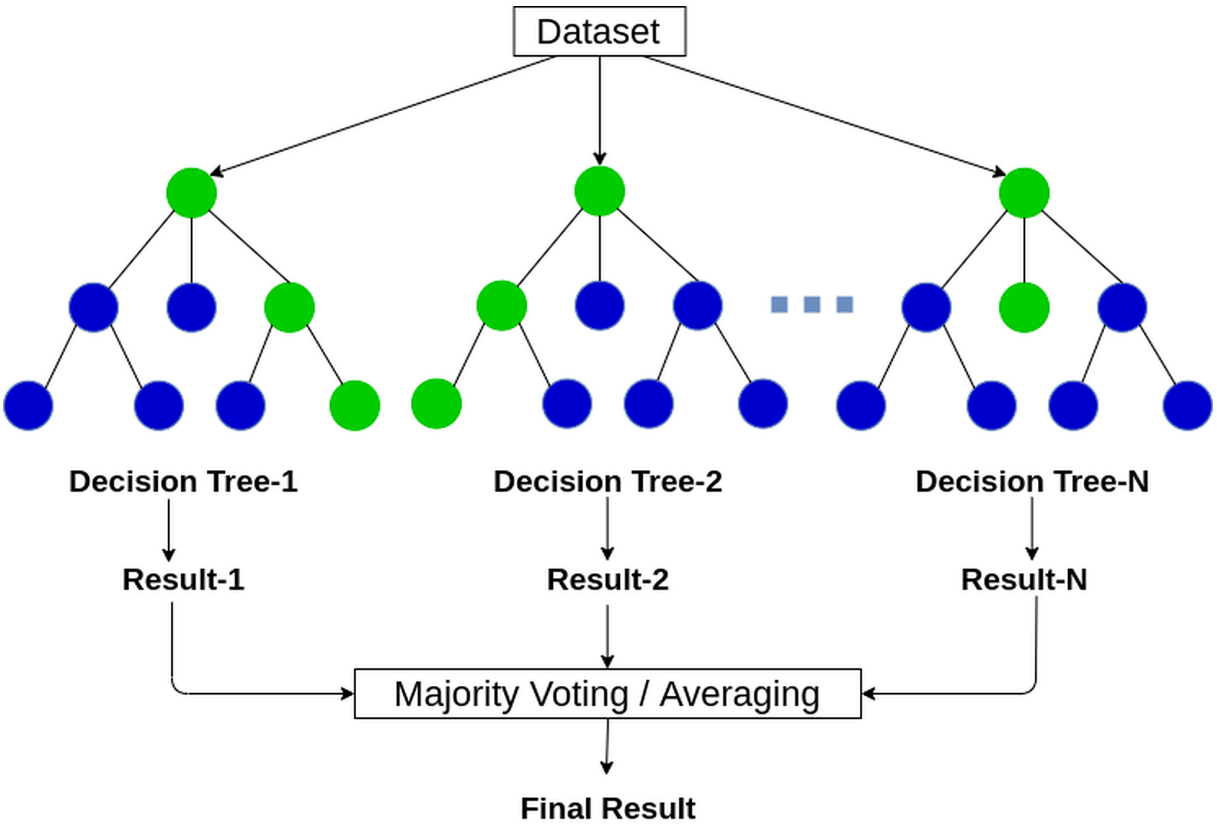


Figure 3.5: Random Forest ensemble learning

Another algorithm that was used in this study is XGBoost [13] which implements gradient boosting decision tree algorithm. XGBoost stands for eXtreme Gradient Boosting. Boosting is an ensemble technique where new models are added sequentially to correct the errors made by existing models until no further improvements can be made [13]. XGBoost trains different models by resampling the data while taking errors made by previously trained models into account. This approach is called gradient boosting where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction [43]. It uses a gradient descent algorithm to minimize the loss when adding new models [43]. This algorithm was considered because of its fast execution performance and better performance.

Tree models can handle data which is not normally distributed; they form visualisable

and interpretable relationships within data and can handle missing data. They apply 'if-then' rules which can easily be interpreted by policy makers because each leaf node is presented as an if-then rule and instances that meet the if-then statement are placed in the node. Model end users prefer prediction models that provide hidden actionable insights that are easy to understand, interpret and validate [45]. In a school context, interpretable machine learning models help school administrators and policy makers to provide proactive feedback and suggest resources to particular schools or students.

3.4.4 Model Hyperparameters

The algorithms used in this study as explained in Section 3.4.1 have variable number of model parameters whose values can be estimated from data. For instance, coefficients in logistic regression and weights in neural networks [28] are examples of internal model parameters that are set from data, but there are some model parameters that can be set manually which are called model hyperparameters. A model hyperparameter is set heuristically to help estimate model parameters which can be used to control the behaviour of the learning algorithms and obtain most skillful model predictions [52].

The core hyperparameters for tree based algorithms included maximum depth of tree, number of leaves in a full tree, minimum gain to split to deal with overfitting and regularization, and bagging fraction to specify the fraction of data to be used for each iteration. The hyperparameters for logistic regression included number of iterations, random state for data shuffling and C for specifying regularization. Think about regularization as any fine-tuning made to a learning algorithm with aim to reduce its generalization error but not its training error. The goal was to train independent Logistic Regression models that generalize better on unseen data without overfitting the training dataset.

3.5 Model Evaluation

The evaluation phase requires evaluating model results and determining possible actions and decisions. In this study, the performance of models was measured using several performance metrics to determine their computational efficiency and accuracy. To improve the accuracy of the models, a stratified 10-folds cross-validator approach was used to train and evaluate our models on every strata. By using the accuracy score, the best performing models were identified to extract important determinants of school outcomes. The performance metrics used to evaluate the performance of models and their stability were: the accuracy score calculated from the confusion matrix, Area Under Curve (AUC), and statistical tests. Goodman and Kruskal's gamma in [17] was used to measure and validate the association between variables, x and y on an ordinal scale, while the Kruskal-Wallistest in [39] was used for x numeric variables against y using the P-value of 0.05.

3.6 Deployment

This is the final phase of CRISP-DM model. Creation of models is usually not the end of a data mining study. Knowledge gained should be organized and presented in a way that stakeholders or end users can use it. Although this study did not produce a deployment plan neither deploy its models, a final report with the results and findings is presented in Chapter 6 in a way that is easily interpreted and useful to end users for education policy and decision making purposes. Building better performing models with high accuracy was not enough. Three approaches were applied to assist in improving the interpretability, usability and effectiveness of the models, namely:

- The Predictive, Descriptive, Relevant (PDR) framework in [47] was used to guide the modelling, extraction and visualization of significant factors influencing school outcomes.

- SHAP (SHapley Additive exPlanations) values [42] and odds ratios explained in [75] were used to interpret models' compositions because of their ability to increase model transparency by indicating how much each predictor contributes.
- Simulation: Simulations were performed to test the effect of a feature availability on the performance of the school. This was done to address the "what if scenarios" for example, what would happen if a school acquired new features such as a library or more teachers.

3.7 Summary

This chapter provided the information about the methods used in this study. It explained the steps that were followed from data collection, analysis and model training. The modelling and analysis processes used in this study was discussed in this chapter. The modelling processes included data preparation, model training and model evaluation. The modelling approach and implementation details of the tools and algorithms used were briefly covered.

Chapter 4

Data

There has been considerable expansion in education data sources. This has improved the availability and certainly the quality of data accessible for research, policy and decision making [81]. Governments and school stakeholders use this education data to extract actionable information for purposes of monitoring and improving schools. This data can be generated in form of big with an immense variety of information collected at increasing volumes and velocity.

This chapter aims to provide the reader with an understanding of the properties of each dataset used in this study. The dataset is described according to the Vs of big data. A summary of the fields is discussed together with data collection and processing tasks. The differences and similarities between the SA and SL datasets are also discussed in this section. The rest of the chapter is organised as follows:

- Section 4.1 presents the definition of big data and a description of how the dataset used in this study fits in the primary 4 Vs of big data.
- Section 3.3 summarises the data preprocessing tasks that were performed to remove the flaws, inconsistencies with the dataset.
- Section 4.2 presents an overview of the South Africa dataset.

- Section 4.3 presents an overview of Sierra Leone dataset.
- Section 4.4 discusses the differences and similarities in the datasets from both countries.
- Section 4.5 provides a summary of this chapter, highlighting notable properties of the datasets.

4.1 Big Data

Big data is data that is too large and too complex to be processed by traditional software. But it comes with many opportunities for analysis, helping schools and governments understand their strengths or weaknesses.

"Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway." – By Geoffrey Moore.

Big data is characterised by 3 primary Vs – Volume, Velocity and Variety listed by Gartner analyst Doug Laney [82]. The fourth important V of big data discussed in this research is Veracity. Datasets used in this research are described according to these four Vs.

Volume refers to the amount of data generated on a daily basis usually measured in bytes. Data used in this research was collected using surveys. It was not large but was collected from over 5000 schools. It also contained over 100 variables were investigated.

Velocity refers to the speed at which data is coming in and getting shared. It is usually measured in real time. For instance social media data is generated every second. Facebook revealed that it collects over 500 terabytes of data each day. Their data comes in at high speed and large volumes. Data used in this research is usually collected once every year.

Variety means big data from multiple sources that is structured, unstructured, and semi

structured. Data used in this research was collected from PDF reports, csv files and GIS locations. One of the criteria used in selecting this data was diversity to ensure that it was properly analysed to come up with insights from multiple sources.

Veracity identifies the credibility of data being generated [65]. Veracity focuses on removing bias, abnormalities and duplicate data with no value. The originality of data used in this research was a critical more than volume and velocity. The data processing Section 3.3 summarises some of the tasks that were performed to ensure that data complied with the veracity element of big data.

4.2 South Africa Datasets

The South African dataset is summarised in Table 4.1. Four main 2016 datasets were used as explained below:

- The high schools' final examination results published by the South African Department of Basic Education. It contains school performance results (pass rates) extracted from the 2016 National Senior Certificate (NSC) school PDF reports.
- The schools master lists. These master lists contain basic information about South African schools such as location, number of teachers and students. They were acquired from the Department of Basic Education website published according to provinces. The master lists were merged with school performance data using unique Education Management Information System (EMIS) school codes.
- The school location data. Although the masterlist contain school geolocation data, most schools had missing longitude and latitude coordinates. The data was collected using Google API services.
- The 2016 community survey was obtained from Statistics South Africa (Stats SA) in a csv format. It contains socioeconomic information about different households

including household facilities such as toilet, radio, television and access to running water. 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.

The final dataset had 31 variables and 5302 (77.8%) secondary schools out of the 6814 schools which sat for 2016 exams across South Africa. It consists of both qualitative and quantitative data. Qualitative data was good for investigating household open reactions such as community water ratings, electricity ratings and hospital ratings. Quantitative education data such as pass rates, and number of teachers and students was good for answering quantitative questions and calculating derived variables.

Note: Schools located in a particular local municipality were assigned the same most prevalent surrounding community and household features. For instance, if in a municipality/community most households had access to clean water and good hospital ratings, all schools in these areas were associated with these facilities. Therefore, schools or students presumed to come from these areas were assumed to have access to similar facilities. This assumption is based on the current school quintile system in SA [19].

Table 4.1: List of variables in the South Africa dataset grouped according to different data data sources: School performance - 2016 exam results, school 2016 master list, and 2016 community survey data

Variable name	Data type	Description
School performance dataset		
EMIS	Integer	Education Management Information System number
Name	String	Name of the school
Province	Category	Province where the school is located
District	Category	District where the school is located
Quintile	Category	Quintile of the school
2016 exam % achieved	Float	Pass rate for 2016 (>= 50 was pass and <50 was fail)
School master list dataset		
EMIS	Integer	Education Management Information System number
Latitiude	Float	GIS latitude
Longititude	Float	GIS longitude
Municipality	Category	Municipality where the school is located
Urban_Rural	Category	Is the school in a rural or urban area
Educators_2016	Integer	Number of teachers
Learners_2016	Integer	Number of Students
Student_Teacher_Ratio	Float	Ratio of students to teachers - calculated variable
2016 Community Survey dataset		
MunicDiff	Category	Difficulties facing the municipality
RateWater	Category	Rating of the overall quality of the water services
RateElectricity	Category	Rating of the overall quality of the electricity supply services
RateToilet	Category	Rating of the overall quality of toilet/sanitation services
RateHospital	Category	Rating of the overall quality of the local public hospital
WaterAccess	Category	Access to safe water supply drinking service
Toilet	Category	Main type of toilet facility used
ToiletLocation	Category	The main toilet facility in the dwelling/yard/outside the yard
MainDwellType	Category	Main dwelling that the household currently lives in
SafetyInDay	Category	Safety during the day
SafetyInDark	Category	Safety when it is dark
ElectrInterrupt	Category	Interruption in electricity in the past 3 months
EnergyLight	Category	Main source of energy for lighting
HeadHH_Age_at_RefNight	Integer	Night Age of household head
HeadHH_Sex	Category	Sex of household head
HHgoods_tv	Category	Household television
Hhgoods_radio	Category	Household radio
HHgoods_dvd	Category	Household DVD/Blu-ray player
Internet_cellphone	Category	Internet-Any place via a cellphone .

4.3 Sierra Leone Datasets

In 2018, the government of Sierra Leone started a school census mission to collect school data to help the government determine the total number of schools by level, type, location, facilities, furniture and enrolment. This was conducted to support inform decision makers in the implementation of the Education Sector Plan 2018-20.

The Sierra Leone dataset summarized in Table 4.2 consists of 3 types of 2018 data sources. Data was acquired from the Directorate of Science, Technology and Innovation and the Ministry of Basic and Senior Secondary Education (MBSSE). The 3 categories of data sources are: final exam results, 2018 school census data and teachers data as explained below:

- The exam datasets contain the 2018 West African Senior School Certificate Examination (WASSCE) average grades for every school.
- The 2018 school survey dataset was carried out in all pre-primary, primary, junior and senior secondary schools. It contains school-level features. In this work, we only extract features of senior secondary schools to map them to WASSCE results.
- The teachers dataset contains details of their sex, teaching experience and qualification, salary source and their schools. This dataset was anonymized and excludes teachers' information such as names and other sensitive information. Teachers were grouped according to schools in order to calculate the number of teachers per school, number of teachers per sex, average teachers' experience in every school and the most teachers' qualification in every school. These calculated variables were merged with school exam and census data using the EMIS codes.

The final dataset has 37 variables and 162 schools in different areas of the country. Through the use of data preprocessing methods presented in Section 3.3, over 50 variables were removed from the dataset because of duplicate values, multicollinearity or having many missing values - where more than 50% of the schools did not provide the information.

Table 4.2: List of variables in Sierra Leone dataset grouped according to data sources

Variable name	Data type	Description
2018 School survey		
emiscode	integer	unique school identification provided by MBSSE
school_name	string	name of the school
latitude	float	school latitude
longitude	float	school longitude
sum_enrol	integer	number of students in the school
remoteness	category	accessibility of the school
idregion	category	region
iddistrict	category	district
school_owner	category	owner of the school
approval_status	category	has the school been approved by the government
financial_support	category	does the school receive financial support
sh_fenced	category	is the school fenced
boarding	category	is it a boarding school
sch_garden	category	does the school have a garden
internet	category	does the school have internet
drink_water	category	does the school have drinking water
computers	integer	number of computers in the school
avail_latrine_fac	category	are toilet facilities available
private_cubicle	category	are there private cubicles
drink_water_source	category	what is the source of the drinking water
library	category	does the school have a library
sci_lab	category	does the school have science lab
canteen	category	does the school have a canteen
rec_facilities	category	does the school have a recreation facility
elec_grid	category	does the school have an electric grid
ssstot	integer	social studies teaching/learning materials
bstot	integer	science teaching/learning materials
mathstot	integer	mathematics teaching/learning materials
basic_comp_skills	category	do the students have basic computer skills
WASSCE exams datasets		
emis_code	integer	unique school identification provided by MBSSE
schName	string	name of the school
papers_passed	float	percentage of papers passed (>= 50 was pass and <50 was fail)
Teachers dataset		
emis_code	integer	unique school identification provided by MBSSE
teacher_time	category	does the teacher work part-time or full time
teacher_prof_qual	category	professional qualification of the teacher
teacher_aca_qual	category	academic qualification of the teacher
teacher_service_years	integer	teacher experience - in years.
teacher_salary_source	category	who pays the teacher (government, private or community)
teacher_sex	category	gender of the teacher

4.4 Differences and similarities in the datasets

This research focused on the impact of school-level and community-level determinants of school performance. It also investigates how performance determinants differ in South Africa and Sierra Leone. However, datasets used are different in size, coverage and content. Variables in the South African dataset mainly contain community-level data and only 5 school-level variables: pass rates, school location, quintile, number of students and teachers. Most variables were extracted from the community household survey. The Sierra Leone dataset mainly consists of school-level features and more teachers' information. Most variables were extracted from the school census survey data. This research used existing datasets and failed to get similar datasets in these 2 countries due to differences in their policies, interests and type of final exams. Contextually, school master lists in SA are similar to the SL annual school surveys, but lack enough details to explain/quantify school facilities and teachers information. The SL dataset also lacks community level details.

4.5 Summary

Education big data is a useful resource in understanding the effectiveness of school environment and determinants of performance. In this chapter 4, several sources of education data including community household survey, school census, teachers information and final examination reports are presented. The Sierra Leone dataset contained more information about schools and teachers whereas the South African dataset contained more information about communities where school are located but lacks information about school facilities and teachers. These datasets were also described according to four primary Vs of big data (Volume, Velocity, Variety and Veracity). Datasets used in this research were mostly found to be characterised by only Variety and Veracity. More granular education datasets should be collected about most school entities to increase the volume of available data and help to solve education issues using a data approach.

Chapter 5

Exploratory Data Analysis

In Chapter 4, the two datasets of interest were introduced. The properties of these datasets were discussed. This chapter presents the Key Performance Indicators (KPIs) set to support decision making about the direction of the work in this chapter. The main goal in this chapter was to obtain confidence in both datasets by extracting meaningful insights to assist in refining feature selection and engaging machine learning algorithms.

The rest of the chapter is organised as follows:

- Section 5.2 presents exploratory insights derived from the SA data.
- Section 5.1 discusses the exploratory insights derived from the SL data.
- Section 5.3 provides a summary of this chapter, highlighting notable insights derived from the SA and SL data.

5.1 Exploring SL Data

The following KPIs were investigated to provide high level familiarity with the Sierra Leone.

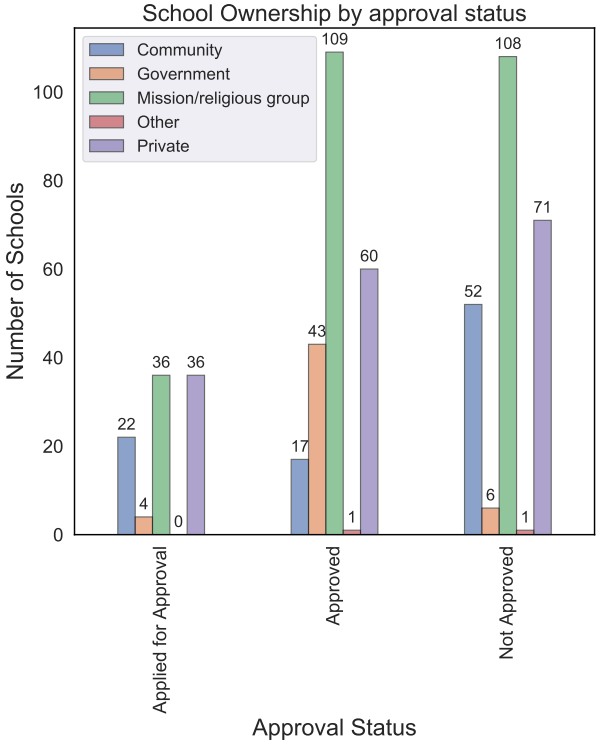
- Number of schools with 2018 WASSCE examination results.
- Number and % of schools by region, district, schools ownership, approval status and financial support.
- Textbook to student ratio by school ownership and district
- Number and % of schools with or without access to facilities such as safe drinking water, latrines, library, electricity and laboratories.
- Number and % of teachers by gender, district and professional qualification.

5.1.1 Schools

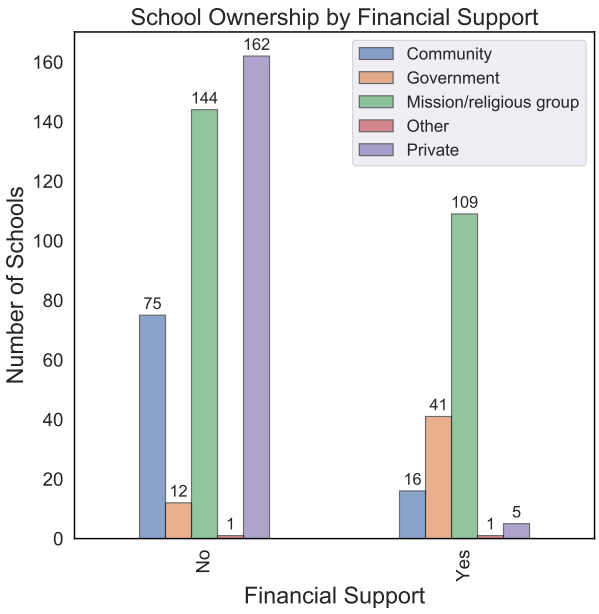
There were 566 senior secondary schools in the Sierra Leone dataset. Only 162 schools had examination results. Exploratory Data Analysis (EDA) results in this chapter explicitly explain features of all the 566 schools. However, 162 schools labelled with exam results were selected for machine learning to train performance prediction models.

Most schools were owned by mission or religious groups (44.7%) and 29.5% of the schools belonged to the private sector. The government only owned 53 of the 566 schools (see table 5.1). It was found that most schools and teachers were located in urban areas. 75.6% of 566 schools were located in urban districts/areas namely; Western area, Bo, Kenema, Kono, Bombali and Port Loko. The Western area comprises the national capital, Bo and Kenema were the second and third largest cities in Sierra Leone.

Figure 5.1: Number of schools by ownership, approval status and financial support



(a) A bar plot showing approval status of various school types. Results showed that most schools were not approved except government schools.



(b) This figure shows number of schools which received financial support. 41 of 53 government schools received and only 5 of 167 private schools received financial support.

5.1.2 Textbooks

Textbooks are learning resources which play an important role in complementing teacher's own resources and experience. The ideal textbook per student ratio is 1:2 or 0.5 (50%) which means 2 students use 1 textbook. In Sierra Leone, most textbook per student ratios in various districts indicated challenges in learning because they were below 50% (see table 5.1). Most schools had more than 10 students sharing a particular textbook (1:10 or or 10% ratio). 58.5% of government schools had no textbooks (ratio = 0).

Table 5.1: Number of schools and textbook to student ratio by school ownership and district. The median of the ratios was considered due to some outliers in the data. The high values of ratios indicate more textbooks available for students to read. Figures in brackets represent the number of schools in each category.

School Owner	Community	Government	Mission/religious group	Other	Private	No. of Schools
District						
BO	0.1(4)	0.1 (5)	0.1 (22)	0 (0)	0.2 (13)	44
BOMBALI	0.1 (12)	0.7 (2)	0.1 (18)	0.2 (1)	0.4 (6)	39
BONTHE	0.8 (1)	0.1 (2)	0.1 (9)	0 (0)	0 (0)	12
FALABA	0.7 (1)	0.1 (2)	0 (1)	0 (0)	0 (0)	4
KAILAHUN	0.4 (1)	0 (0)	0 (17)	0 (0)	0 (0)	18
KAMBIA	0.2 (9)	0.2 (3)	0.1 (14)	0 (0)	0 (0)	26
KARENE	0.2 (5)	0 (0)	0.3 (6)	0 (0)	0 (0)	11
KENEMA	0.3 (5)	0 (3)	0 (21)	0 (0)	0.1 (15)	44
KOINADUGU	0.2 (2)	0 (1)	0 (5)	0 (0)	2 (1)	9
KONO	0.1 (11)	0 (2)	0.1 (21)	0 (0)	0.5 (4)	38
MOYAMBA	0.1 (2)	0.2 (2)	0.1 (18)	0 (0)	1.3 (4)	26
PORT LOKO	0 (7)	0 (4)	0.1 (25)	0 (0)	0.1 (4)	40
PUJEHUN	0 (0)	0.5 (1)	0.3 (6)	0 (0)	0 (0)	7
TONKOLILI	0.1 (4)	0.1 (5)	0.1 (15)	0 (0)	0 (1)	25
WESTERN AREA RURAL	0.2 (14)	0 (2)	0.1 (26)	0 (0)	0.2 (33)	75
WESTERN AREA URBAN	0.2 (13)	0 (19)	0.1 (29)	0 (1)	0.2 (86)	148
No. of Schools	**91**	**53**	**253**	**2**	**167**	**566**

5.1.3 Summary of numeric variables

The Sierra Leone dataset had many continuous variables which this chapter analyzed to check correlations and multicollinearity among them. Figure 5.2

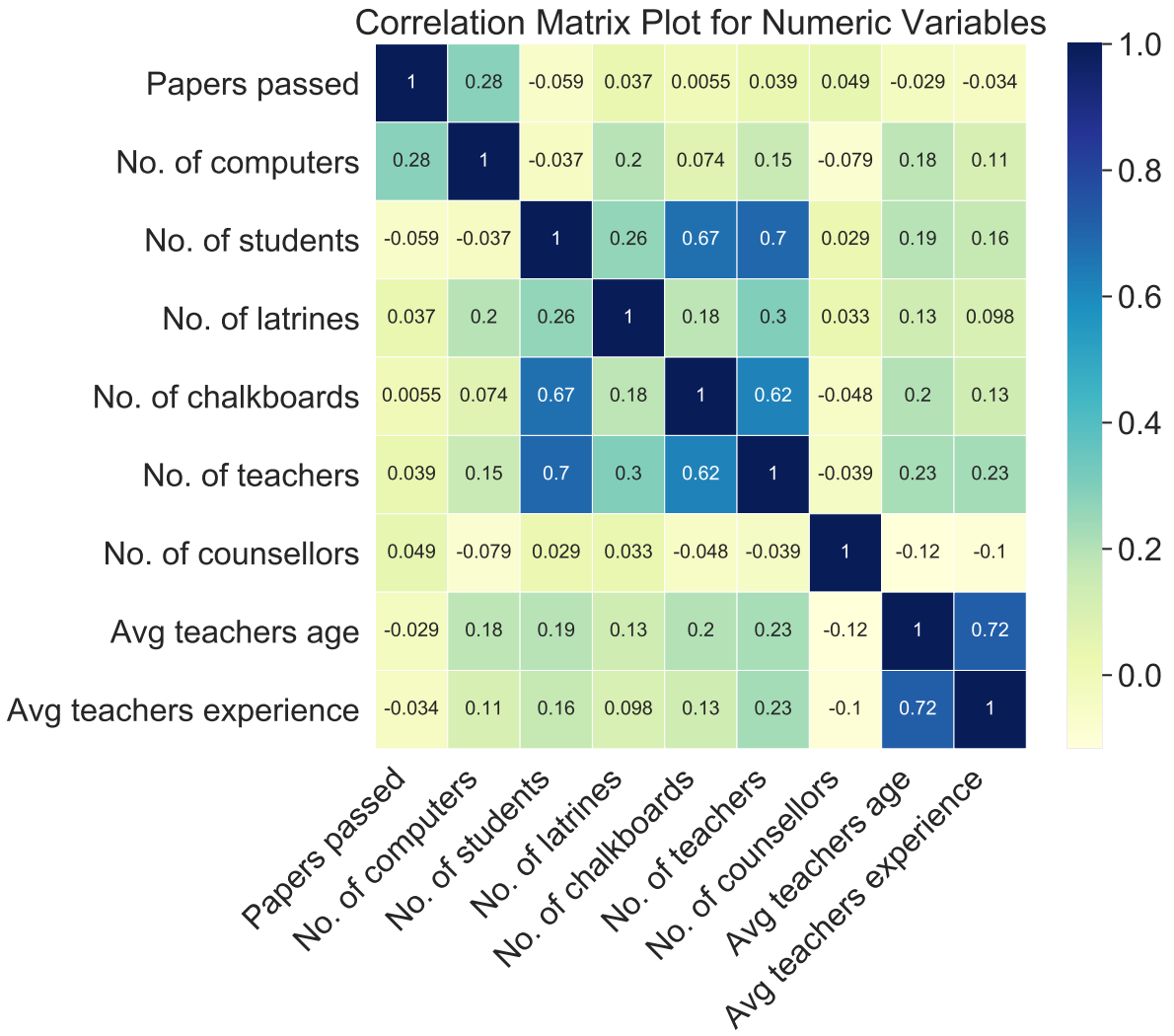


Figure 5.2: A correlation matrix heatmap showing correlations among continuous variables for Sierra Leone dataset. The plot is annotated with Pearson's correlation coefficients for $P - Value < 0.05$. Blue means positive, yellow means negative. There was a high positive correlation between number of teachers and students (0.7). This mainly schools with more students had more teachers and vice versa. Most features were almost uncorrelated with the average number of papers passed in schools.

5.1.4 Prevalence of school facilities with performance

This section presents a summary of school facilities and the number of schools which belonged to every performance category. Schools were categorised into 4 groups based on the quantiles of papers passed and their responses. All schools in the first and second quantile (1Q and 2Q) failed - with their range of papers passed as [6.4, 31.3] and (31.3, 47.4] respectively. Some schools in 3Q failed. 4Q consisted of the top schools where most students passed more than 68.6% of the papers.

Table 5.2: A frequency table showing the number of schools in every performance category based on their facilities

	1Q	2Q	3Q		4Q
	[6.4, 31.3]	(31.3, 47.4]	(47.4, 68.6]	(47.4, 68.6]	(68.6, 99.6]
Variable	**fail**	**fail**	**fail**	**pass**	**pass**
remoteness: accessible	39	40	3	32	37
remoteness: rough terrains	2	0	1	2	4
school owner: community	4	5	0	3	1
school owner: government	9	9	0	8	8
school owner: mission group	26	24	4	22	17
school owner: private	2	2	0	3	15
boarding: yes	3	5	1	7	5
boarding: no	38	35	3	29	36
drinking water: yes	40	34	4	34	35
drinking water: no	1	6	0	2	6
library: yes	28	21	3	24	29
library: no	13	19	1	12	12
canteen: yes	5	4	1	10	19
canteen: no	36	36	3	26	22
electricity grid: yes	20	22	1	18	34
electricity grid: no	21	18	3	18	7
fence: yes	18	21	1	17	32
fence: no	23	19	3	19	9
internet: yes	4	4	0	5	7
internet: no	5	9	2	4	13
internet: unknown	32	27	2	27	21
available latrine: yes	41	40	4	33	40
available latrine: no	0	0	0	3	1
public cubicle: yes	4	6	0	4	9
public cubicle: no	37	34	4	32	32
science lab: yes	20	26	3	21	24
science lab: no	21	14	1	15	17
recreation facility: yes	35	33	3	30	30
recreation facility: no	6	7	1	6	11
generator: yes	20	15	2	16	23
generator: no	21	25	2	20	18
basic computer skills: yes	8	12	2	8	16
basic computer skills: no	33	28	2	28	25
financial support: yes	35	33	3	29	24
financial support: no	6	7	1	7	17

5.1.5 Teachers and students

Teachers are essential in educating and mentoring students. They also play an important role of building conducive learning environments for sharing knowledge with students. Exploratory results in this section report high level information about teachers variables found in the Sierra Leone dataset. There were 10116 senior secondary teachers. Male teachers dominated the teaching industry raising a percentage of 91.3 of the total number of teachers. 78.3% of the teachers were located in the top urban cities which had most schools. It was found that 73.8% of the teachers had professional qualifications while 26.2% (2654 teachers) did not receive formal training as educators but had other degrees and qualifications.

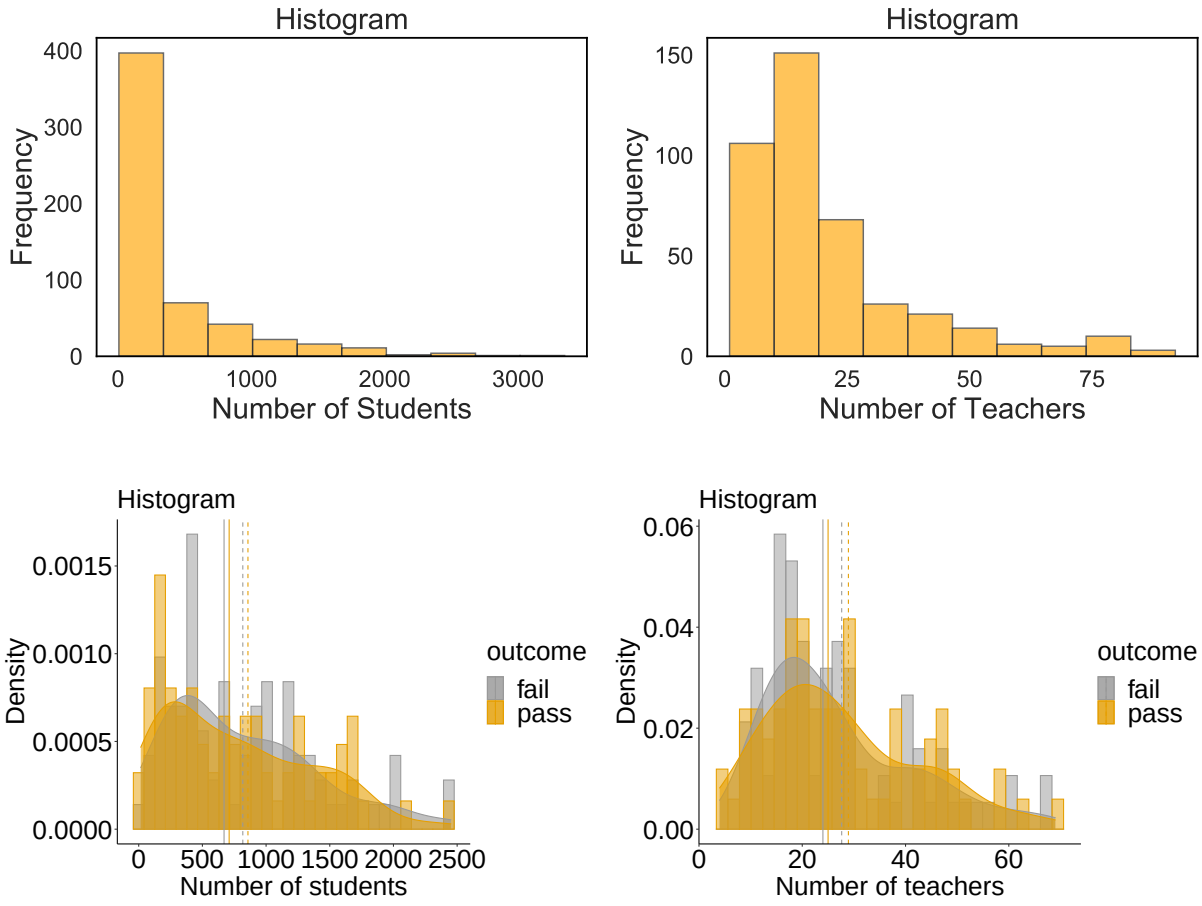


Figure 5.3: Histograms showing number of teacher and students in Sierra Leone schools. The top histograms represent values for the 566 schools. Bottom histograms represent the distribution of values for only 162 schools which had examination results. For bottom histograms, dashed lines indicate the mean in each category while the solid line indicate median values.

The overall average number of teachers in schools was 21 (median = 15, SD = 17). 75% of the schools did not have more than 25 teachers. 75% of the schools had their student population falling under 425. The median number of senior students in schools was 165 (mean 353, SD = 468). The teacher per student ratio in 75% of the schools was 11:100 (mean = 7:50 and median = 3:50). However, only 162 of 566 schools had examinations results in acquired dataset. The descriptive statistics of the sampled 162 schools varied from the overall dataset statistics. For instance, the average number of students in senior

schools with exam results was 819 (median = 668 and SD = 640). 75% of 162 schools had their student population falling under 1216. The average number of teachers was 29 (median = 24). There was no significant difference in the average number of teachers and students in both pass and fail school categories as shown by the lines in Figure 5.3. For example, the average number of students in schools which failed was 817 (median = 670) and 857 (median = 710) for schools which passed. The median number of teachers in schools which passed was 25 and 24 for schools in the fail category.

5.2 Exploring The SA data

In order to understand the nature of the South African dataset, the EDA was performed using the KPIs below to identify the properties of the dataset to determine which features were most appropriate for machine learning algorithms. The analysis was conducted to investigate the spread among the members of the data, skewness of the data, outliers and correlations among the elements in the dataset. The following KPIs guided the analysis to provide familiarity and build trust with the South Africa dataset.

- Number of schools and their performance by province.
- Performance of schools according to quintiles.
- Features of best performing schools (with 100% pass rate).
- Performance of schools in areas with good or poor community facility ratings.
- Teacher student ratio.

5.2.1 Performance of schools in various provinces

There were 5302 schools in the South African dataset. The top performing provinces were Free State (100%), Gauteng (98.9%), Western Cape (98.3%) and North West (96.3%).

Eastern Cape and Limpopo were the least performers with 38.7% and 32.0% of their schools failing to raise above average pass rate respectively. These provincial performance rankings were similar to those published in the National Senior Certificate school performance report of 2016.

The performance of schools depended on the quintile number of the school (chi square = 480.1, P-Value = $1.4e^{-102}$). The percentage of schools which passed in every quintile category was as follows: 99.6% for quintile 5, 97.6% for quintile 4, 81.8% for quintile 3, 76.3% for quintile 2 and 69.5% for quintile 1 schools. Lower quintiles indicated high likelihood of schools failing to achieve more than 50% pass rate.

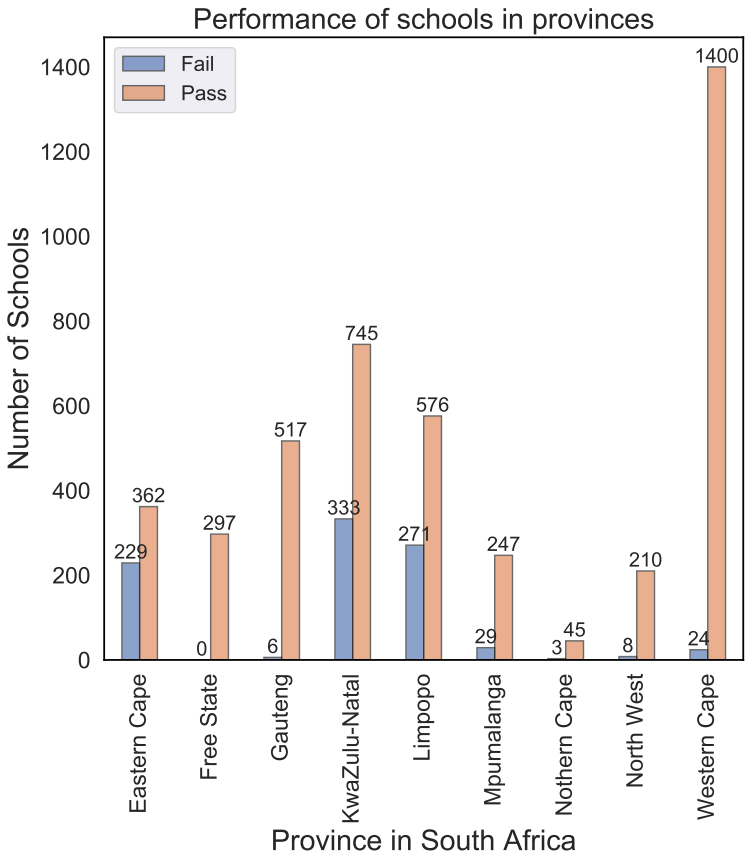


Figure 5.4: Bar chart showing number of schools which passed or failed in every province. Fail means the school's pass rate was less than 50% while schools in the category of pass managed to get more than or 50% of their students pass the final examination.

5.2.2 Performance of schools according to the most prevalent facilities in communities

Electricity, water and hospitals are mostly shared facilities in communities. Water shortages or electricity blackouts mostly affect the whole community. The EDA process identified the number of schools with their performance category (pass or fail) in communities with different facilities and their ratings. Results in Table 5.3 the number of schools in every community facility category. Schools were divided into 3 groups: strong, fair and weak schools. Strong schools (404 of 5302) achieved 100% pass rate. Schools which scored less or equal to 40% schools were labelled as struggling schools, namely, fair schools (with 21% to 40% pass rate) and weak schools with less than 21% pass rate.

The following insights were observed about these schools:

- 64.1% of the strong schools were quintile 5 (259) schools and 14.1% were quintile 4 (57) schools.
- 85.9% of strong schools were located in urban areas compared to only 22.0% and 12.6% of fair and weak schools respectively.
- The prevalence of most important features reduced with performance which indicated that only a few weak or fair schools had access to these facilities compared to strong schools. Most strong schools were situated in areas with good facility ratings such as water, electricity, hospitals and toilets. Fair schools followed strong ones in raising higher percentages of schools situated in areas with good facility ratings and lastly weak schools. For instance, 91.1% of strong schools were in areas with good hospitals but 77.6% of fair schools were situated in similar areas and the percentage further reduced to 66.4% of weak schools in communities with good hospitals.

Table 5.3: Frequency of categorical features in South African communities with the strong 404 schools which scored 100%, and 565 struggling schools with <=40% pass rate separated into groups: 119 weak schools (with 0-20% pass rate) and 446 fair schools (with 21-40% pass rate). There was no strong school located in a community with no electricity, while 1 weak school (0.8% of 199) and 4 fair schools (0.9% of 446) were located in areas with no electricity.

	Strong		Fair		Weak	
Variable	**number**	**percentage**	**number**	**percentage**	**number**	**percentage**
Quitile: 5	259	64.1	2	0.4	0	0.0
Quitile: 4	57	14.1	3	0.7	1	0.8
Quitile: 3	22	5.4	104	23.3	22	18.5
Quitile: 2	36	8.9	146	32.7	29	24.4
Quitile: 1	30	7.4	191	42.8	67	56.3
Urban_Rural: Urban	347	85.9	98	22.0	15	12.6
Urban_Rural: Rural	57	14.1	348	78.0	104	87.4
RateHospital: good	368	91.1	346	77.6	79	66.4
RateHospital: average	20	5.0	96	21.5	40	33.6
RateHospital: poor	12	3.0	1	0.2	0	0.0
RateWater: good	349	86.4	266	59.6	49	41.2
RateWater: average	31	7.7	34	7.6	22	18.5
RateWater: poor	24	5.9	146	32.7	48	40.3
WaterAccess: yes	401	99.3	378	84.8	98	82.4
WaterAccess: no	3	0.7	68	15.2	21	17.6
RateElectricity: good	399	98.8	433	97.1	114	95.8
RateElectricity: average	5	1.2	9	2.0	4	3.4
ElectrInterrupt: yes	3	0.7	10	2.2	3	2.5
ElectrInterrupt: no	401	99.3	436	97.8	116	97.5
EnergyLight: electricity	404	100.0	438	98.2	118	99.2
EnergyLight: candles	0	0.0	8	1.8	1	0.8
RateToilet: good	395	97.8	385	86.3	97	81.5
RateToilet: average	8	2.0	46	10.3	17	14.3
RateToilet: poor	1	0.2	13	2.9	5	4.2
SafetyInDark: safe	64	15.8	64	14.3	27	22.7
SafetyInDark: unsafe	340	84.2	382	85.7	92	77.3
SafeInDay: very-safe	305	75.5	408	91.5	109	91.6
SafeInDay: fairly-safe	99	24.5	38	8.5	10	8.4
HHgoods_radio: yes	399	98.8	406	91.0	105	88.2
Hhgoods_radio: no	5	1.2	40	9.0	14	11.8
Internet_cellphone: yes	189	46.8	65	14.6	10	8.4
Internet_cellphone: no	215	53.2	381	85.4	109	91.6
HHgoods_dvd: yes	343	84.9	149	33.4	20	16.8
HHgoods_dvd: no	61	15.1	297	66.6	99	83.2
MainDwellType: formal	393	97.3	319	71.5	74	62.2
MainDwellType: traditional	11	2.7	127	28.5	45	37.8
HHgoods_tv: yes	401	99.3	434	97.3	115	96.6
Hhgoods_tv: no	3	0.7	12	2.7	4	3.4

5.2.3 Teachers and students

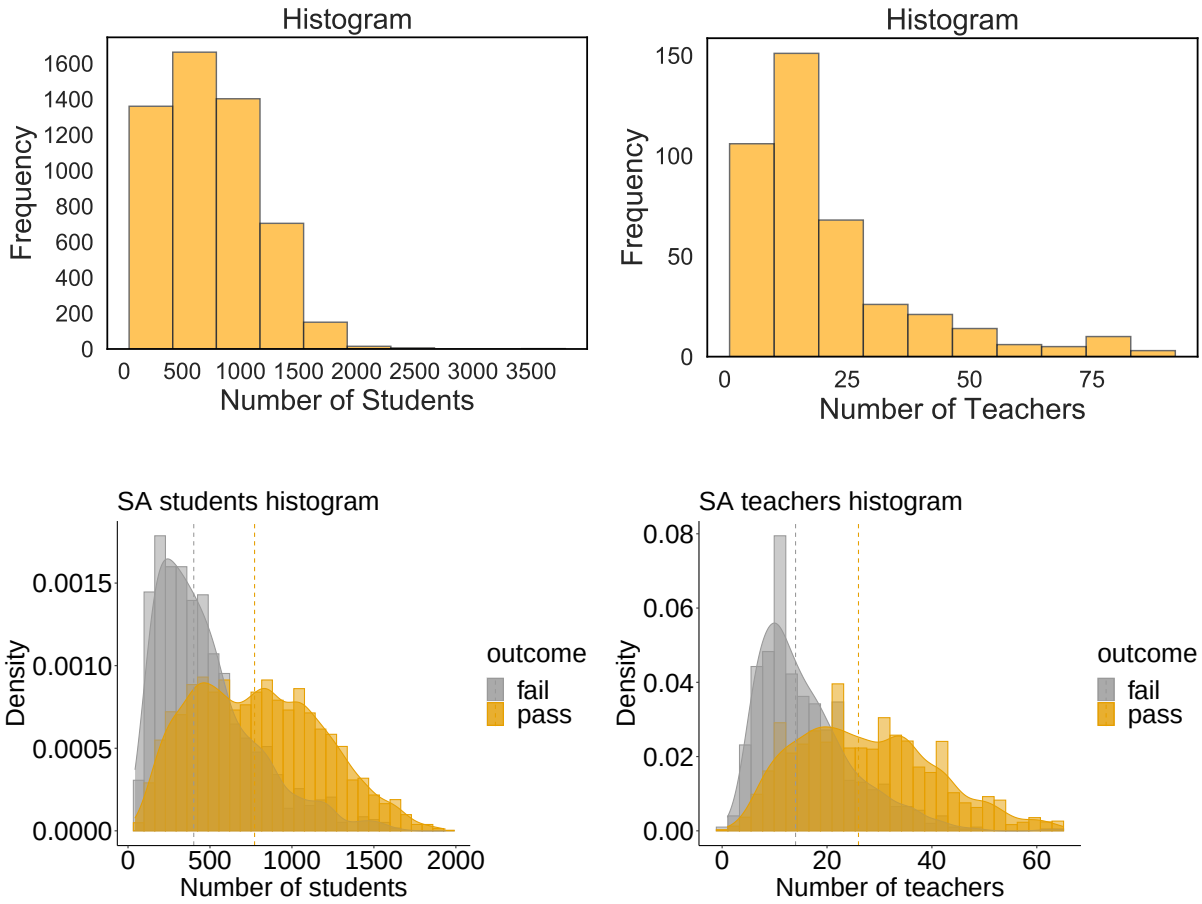


Figure 5.5: Histogram showing number of teachers and students for SA. Dashed lines indicate the mean in each category while the solid line indicate the median values. The overall median number of students 692 (mean = 731, 75th percentile = 1022) and the number of students in schools which passed vs failed was 775 and 402 respectively. The overall median number of teachers was 24 (75th percentile = 35). A significant difference in the number of teachers was identified where schools which passed had 27 teachers (median) compared to the 14 teachers of fail schools but the overall average teacher per student ratio was the same for both school categories (3:100)

5.2.4 Difference of performance outcomes in both countries

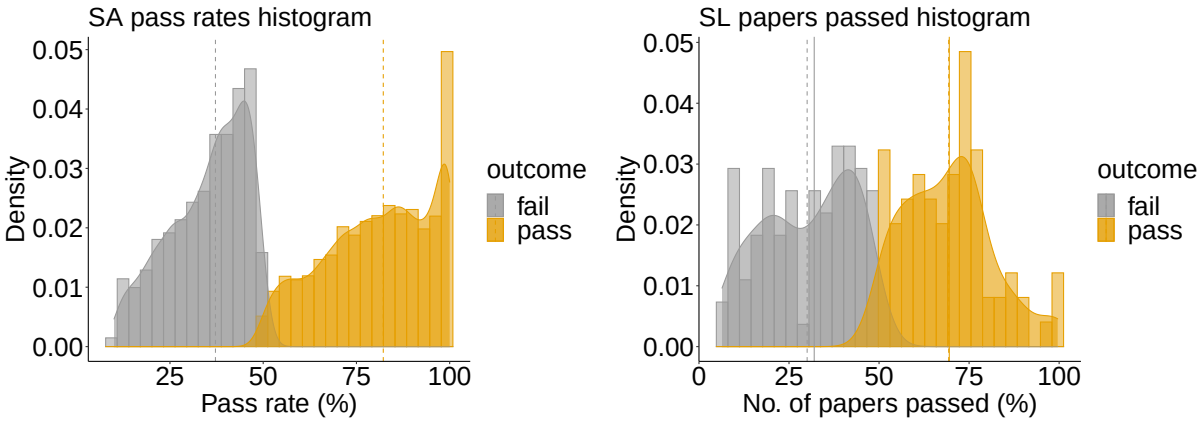


Figure 5.6: Histograms showing distributions of school outcomes for both countries. The performance of schools in SA was measured by pass rates as shown on the left while that of SL schools was measured in terms of number of papers passed by students as shown on the right. Dashed lines indicate the mean in each category while the solid line indicate the median values. For SA, the average pass rate in schools which failed (in grey) was 34.2 (median = 36.9) compared to 80.7 (median = 82.3) in schools which passed. For SL, the average number of papers passed for schools in the fail category was 30% (median = 32) while schools which passed attained 69.3% (median = 69.5).

5.3 Summary

The exploratory analysis conducted on the datasets focused on providing descriptive statistics of the variables, identifying Outliers, correlations among predictor variables with measures of performance. Frequency tables were used to identify the prevalence of school and community facilities in areas where schools performed well or poor.

The exploratory analysis uncovered that quintile 5 and 4 schools perform better than other quintile schools in South Africa. The best performing provinces in South Africa were Free State, Gauteng and Western Cape. Schools in top Sierra Leone cities also

performed better than those in rural districts. Features for Only 162 Sierra Leone schools with examination data were selected to train machine learning models.

The insights uncovered in the exploratory analysis provides a basis for further analysis and discussions about determinants of school performance in senior secondary final examinations. The next chapter provides insights into the predictive results obtained from machine learning experiments.

Chapter 6

Predictive Modelling

The content of this chapter provides an account of every model performance. The interpretation of model results and how they differ in every country.

The remainder of this chapter is organised as follows:

- Section 6.1 provides information about the performance of trained model measured using commonly used metrics. The values of parameters used to control the learning process of the models are also presented in Table 6.2.
- Section 6.2 discusses the most important performance determinants from the South African data.
- Section 6.3 discusses significant features ranked by the models trained on the Sierra Leone data.
- Section 6.4 summarises the contents of this chapter.

6.1 Modelling Results

The machine learning algorithms used to train models whose results are presented in this chapter were discussed in the methodology chapter under section 3.4.1. Their mode of execution was also discussed together with their theoretical performance variations. These algorithms included Logistic Regression and tree based algorithms such as XGBoost, Random forest and decision trees.

This section's objective is to present models which achieved superior performance (e.g. accuracy) on the test data. Models' performances varied across countries since they were each trained using different datasets and variables. The training datasets had balanced number of predicted classes, which is why accuracy was one of the top performance metrics. It is also easily understood by non-technical stakeholders.

Table 6.1 shows test-set average performance scores of machine learning models trained and tested in 10 experiments. Experiments were conducted using various hyperparameter search trials in order to optimize each model performance. Table 6.2 presents values of the types of parameters that were used to control the learning process of the models. Some of the parameters changed because the datasets were differed in size.

These performance results are not sufficient for drawing accurate conclusions about which features were more significant and associated to various schools' performance, but Sections 6.2 and 6.3 discuss additional details of interpretable machine learning techniques which were used to extract the most important predictive features.

XGBoost outperformed other classifiers with the accuracy of 65.5% and 75.5% , and the Area Under Curve (AUC) of 77.3% and 82.1% for SL and SA respectively. XGBoost is an optimized distributed gradient boosting algorithm which can have better performance than RF and DT if parameters are correctly tuned. It also acquired the highest specificity on both datasets. Specificity means the ability of the model to correctly identify good schools with pass rates or percentage of papers passed greater or equal to 50%. The XGBoost model was ranked third in correctly identifying SL failing schools with 67.7% performance after RF with 71.1% and LR which led with 73.3%.

The AUC was used to choose the best tree-based model (XGBoost) for feature extraction purposes. It is a good measure of separability since it can show how much a model is capable of distinguishing between classes. The higher the AUC, the better the model. The LR model which scored the second best AUC in SA and third best AUC in SL was also selected for feature ranking to extracting the odds ratios and validate XGBoost top features on a linear scale.

Table 6.1: Average performance of models (in %) on South Africa and Sierra Leone datasets.

	South Africa				Sierra Leone			
Model	**Accuracy**	**Sentivity**	**Specificity**	**AUC**	**Accuracy**	**Sentivity**	**Specificity**	**AUC**
XGBoost	75.5	67.7	82.9	82.1	65.5	63.2	78.5	77.3
Logistic regression	72.3	73.3	71.4	79.5	61.2	48.1	77.2	60.3
Decision trees	75.1	59	90.6	79.1	61.2	59.3	63.6	63.1
Random forests	70.5	71.1	69.9	78.8	55.1	44.4	68.2	55.7

Table 6.2: Hyperparameters

Algorithm	**Parameter**	**SL - values**	**SA - values**	**Description**
XGBoost	scale_pos_weight	1	1	Control the balance of positive and negative weights
	learning_rate	0.01	0.01	Step size shrinkage used in update to prevents overfitting
	colsample_bytree	0.3	0.4	pecify the fraction of columns to be subsampled
	subsample	0.8	0.3	Subsample ratio of the training instance
	objective	binary:logistic	binary:logistic	logistic regression for binary classification, output probability
	reg_alpha	0.3	0.3	Increasing this value will make model more conservative
	max_depth	8	3	Maximum depth of a tree
	gamma	0	1	Minimum loss reduction required to partition the leaf node of the tree.
Random Forest	bootstrap	TRUE	TRUE	Whether bootstrap samples are used when building trees
	ccp_alpha	0	0	used for Minimal Cost-Complexity Pruning. No prunning was performed
	min_samples_leaf	1	1	The minimum number of samples required to be at a leaf node
	min_samples_split	2	2	The minimum number of samples required to split an internal node
	n_estimators	10	100	The number of trees in the forest.
	criterion	entropy	entropy	The function to measure the quality of a split
Decision Tree	criterion	gini	gini	The function to measure the quality of a split
	max_depth	8	20	The maximum depth of the tree
	min_samples_leaf	1	1	The minimum number of samples required to split an internal node
	class_weight	balanced	balanced	Weights associated with classes
	random_state	22	42	Controls the randomness of the estimator
Logistic Regression	solver	liblinear	liblinear	Algorithm to use in the optimization problem
	penalty	l2	l2	Used to specify the norm used in the penalization
	C	1	1	Inverse of regularization strength

6.2 Interpretation of South Africa results

School quintiles, location (rural or urban), good hospitals, access to good water and availability of DVDs and cell phone Internet were found to be significant in predicting good performance in South Africa.Poor toilets, electricity interruptions, traditional dwellings, had higher importance in predicting the fail outcome in South Africa.

Table 5.3 shows the prevalence (frequency) of performance determinants in schools which scored 100% (404 strong schools) and in schools with less than or equal to 40% pass rates (565 struggling schools). Struggling schools were further divided in two groups namely; weak schools (with 0-20% pass rate) and fair schools (with 21-40% pass rate). There were 119 weak schools and 446 fair schools raising a percentage of 2.2% and 8.4% of the whole dataset respectively, while 7.6% were strong schools.

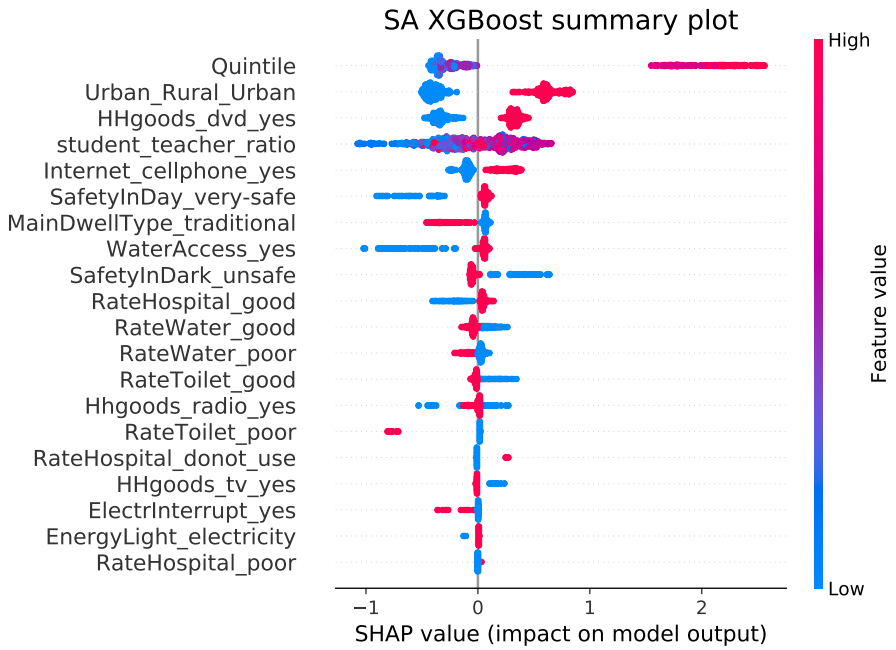
Schools in Urban areas performed better than rural areas which indicated that the location of the school had a high positive impact on performance ($\gamma = 0.74$ at P-value $< 2.2e^{-16}$). This variable was ranked first and second by LR and XGBoost models respectively (see Figure 6.1).

Table 6.3: SA Logistic regression odd ratios. For a school in an urban area, the odds of pass vs. fail were by a factor of 3.86 or would increase by 285.74%

Variables	**weights**	**odd-ratio**	**%Change**
RateWater: good	-0.93	0.40	-60.42
RateWater: poor	-0.60	0.55	-45.30
RateToilet: good	-0.53	0.59	-41.23
RateToilet: no-access	-0.39	0.67	-32.53
RateToilet: poor	-1.85	0.16	-84.32
Urban_Rural: urban	1.35	3.86	285.74
RateHospital: donot-use	0.61	1.85	84.63
RateHospital: good	0.35	1.42	41.51
RateHospital: poor	0.11	1.12	12.12
WaterAccess: yes	0.73	2.08	107.53
MainDwellType: traditional	-0.69	0.50	-49.66
SafetyInDay: very-safe	1.23	3.44	243.75
SafetyInDark: unsafe	-0.76	0.47	-53.36
ElectrInterrupt: yes	0.41	1.51	51.29
EnergyLight: electricity	0.14	1.15	15.49
HHgoods_tv: yes	-1.35	0.26	-74.14
HHgoods_radio: yes	0.02	1.02	1.66
HHgoods_dvd: yes	0.85	2.35	134.61
Internet_cellphone: yes	0.45	1.57	57.45
Quintile	0.57	1.77	76.86
student-teacher ratio	0.004	1.004	0.415

Figure 6.1:

(a) A bar chart showing the impact of every feature in the logistic regression model. Urban schools, very safe safety in the day, access to water, quintiles and other features whose bars are on the right increased the odds of passing while poor toilets, traditional dwellings and insecurity at night reduced the odds of passing.



(b) A summary plot showing the distribution of every feature impact in the XGBoost model. The color represents the feature value (red represents high and blue represents low). Low quintiles lowered performance in schools, high cases of cellphone internet usage increased performance and high cases of traditional dwellings were associated to decrease in the performance of students in schools.

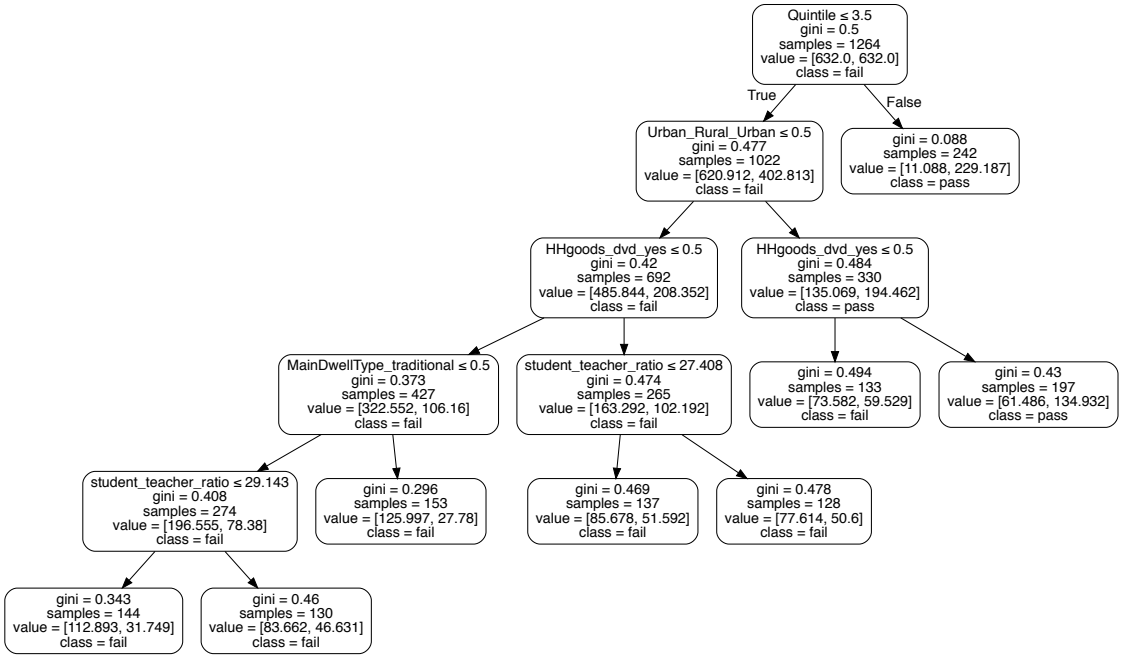


Figure 6.2: A pruned decision tree for South Africa: The split was formed on quintiles followed by the rural-urban school location divides. Schools with quintile 4 and 5 were categorised in the pass category with a Gini impurity of 0.088 whereas urban schools located in areas where most households had DVDs were also categorised in the pass category.

6.2.1 School quintiles and community infrastructure

The quintile ranking system of public ordinary schools in South Africa was found to have a high positive impact on performance and the decision tree in Figure 6.2 considered quintiles at the core of the splitting phase. Schools in quintile 5 and 4 performed better than those in lower quintiles. Results show that 64.1% of the strong schools belonged to quintile 5 and there was no weak school in the quintile 5 class. 56.3% of weak schools belonged to quintile 1.

There are 5 categories of quintiles, largely for purposes of allocating financial resources. The poorest quintile is 1 and least poor is 5. These poverty rankings are determined nationally according to the poverty of the community around the school, as well as,

certain infrastructural factors. These infrastructural factors can be electricity, hospitals, roads and water. The following results explained how these community infrastructure developments were founded to be associated with the performance of schools located in these communities.

Good hospitals were found to be significant with its unit increase contributing 41.5% change to the odds of pass vs fail. The XGBoost model associated lower hospital ratings with a negative impact on performance outcomes (γ = 0.42; P-value $< 1.2e^{-15}$). It was further found that 91.1% of strong schools were located in communities with good hospital rating compared to 77.6% of fair and 66.4% of weak schools.

Schools located in areas with mostly traditional dwellings and poor toilet ratings did not perform well. These features was found to reduce the odd ratios of pass vs fail in schools.

Electricity ratings were good in most communities and over 95% of South African schools were located in areas with good electricity ratings. However, results found electricity interruptions explaining the variations in pass and fail. Higher values of electricity interruptions had a negative impact on school performance (see Figure 6.1b; γ = -0.41; CI: (-0.65, -0.16) ; p-value = 0.02).

Access to water in communities was a good predictor of school outcome (γ = 0.68; CI = (0.61 0.75) and found to increase the odds of pass vs fail by a factor of 2.08 given all other features stay the same. The XGBoost model also associated very low cases of water access in communities to lower school outcomes.

6.3 Interpretation of Sierra Leone results

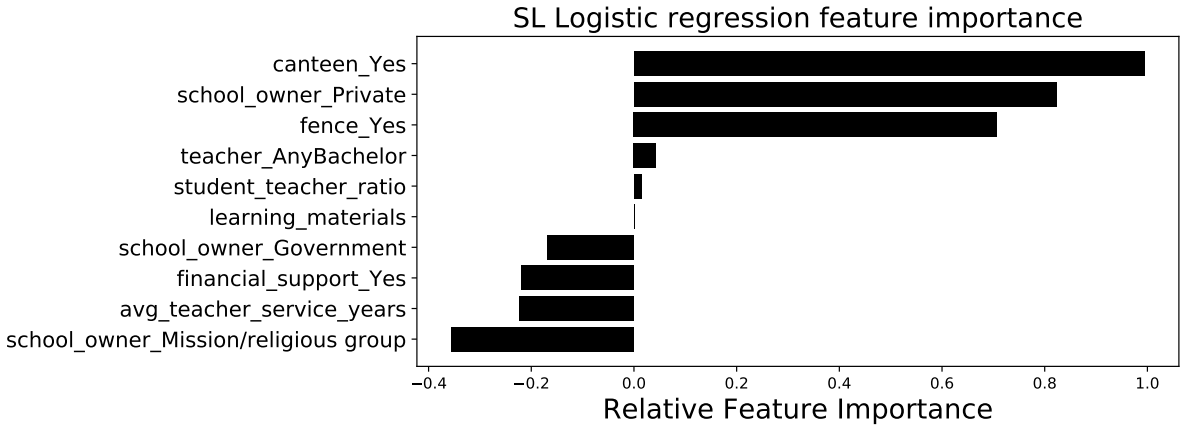
Results show that private schools, school location, availability of canteen, fence, electricity, and number of teachers with bachelor degrees were the most significant features in predicting the pass outcome. Availability of school fence was associated with more

cases of passing in schools as 55% of fenced schools passed compared to the 38% unfenced schools. Schools in the Western region (urban region) performed better than other regions. Mission and government school ownership, financial support were found with higher importance in predicting the fail outcome as shown in Figure 6.3.

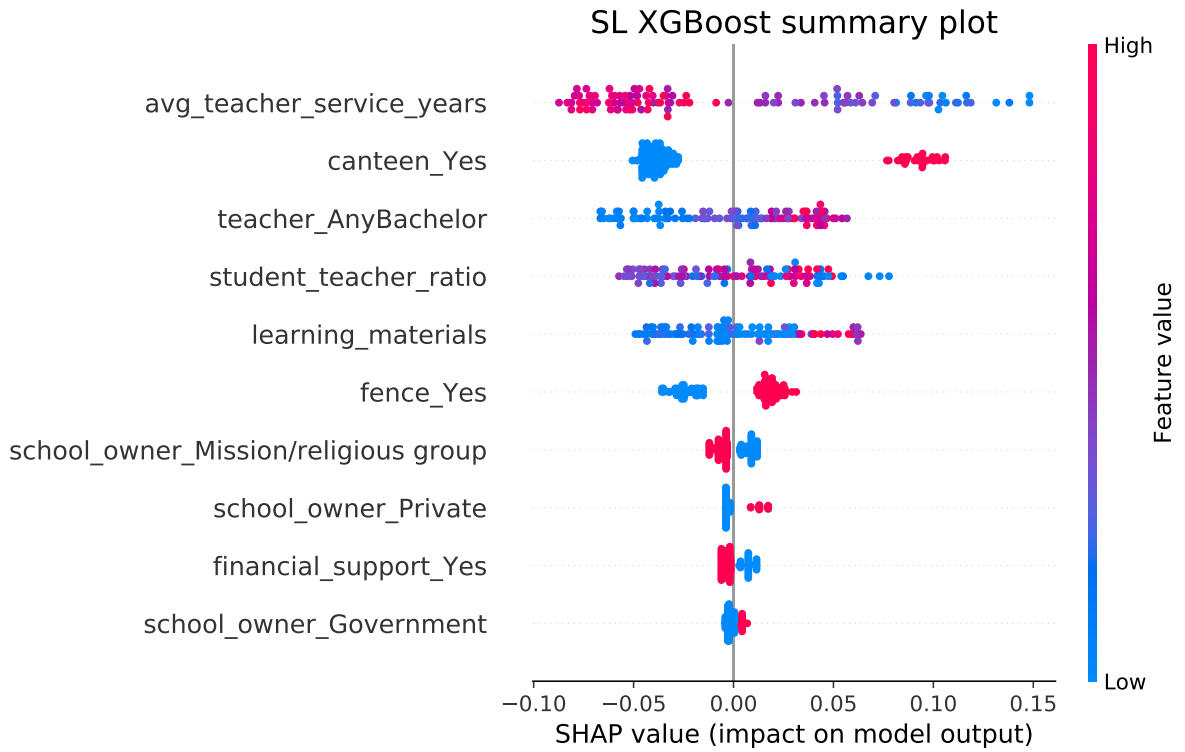
6.3.1 School Canteens

Availability of canteens commonly known as restaurants in schools was ranked at the top by both LR and XGBoost models in having a positive impact on the school outcomes as shown in Figure 6.3 and 6.4. Canteens provide feeding as a social safety net with its other benefits like reducing school drop-outs discussed in [30]. Results showed that 74.4% of schools which had canteen passed while only 38.3% of schools with no canteens passed.

Figure 6.3:



(a) This bar chart was extracted from the LR model and it represents Sierra Leone feature rankings. Availability of canteen, school fence and private school status increased the odds of passing in schools while government or mission schools with more financial support reduced the odds of passing.



(b) A summary plot showing the distribution of Sierra Leone features as extracted from the XGBoost model. High values are represented by the red colour and low values are in blue. High values of average teaching experience were associated to low performance and schools with no fences were associated with low students' performance.

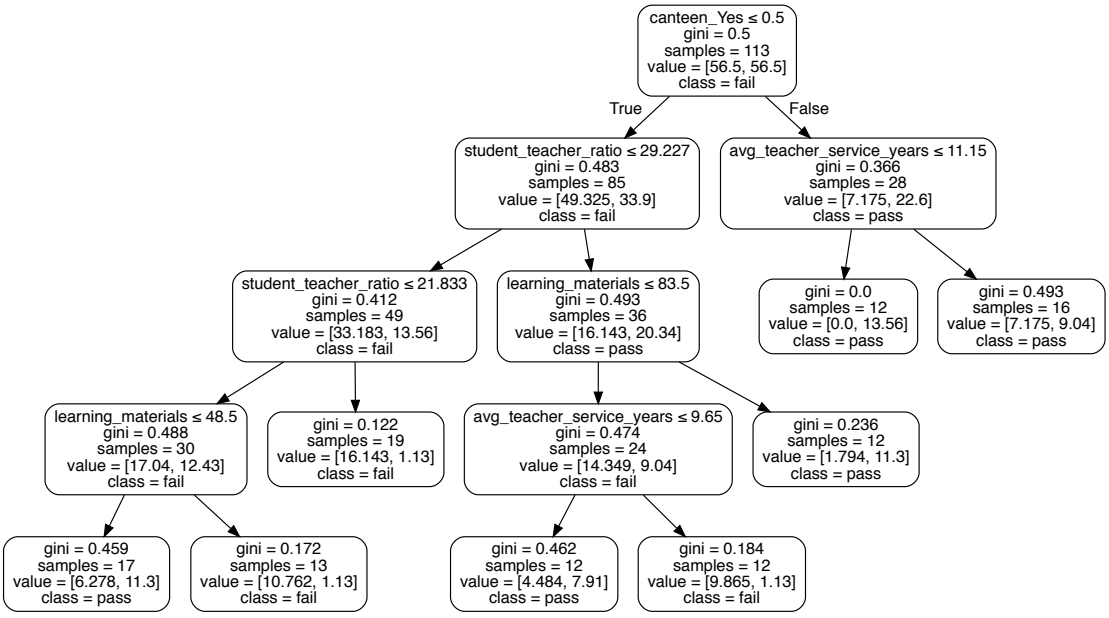


Figure 6.4: Decision tree for Sierra Leone: Like XGBoost and LR models, the availability of canteens in Sierra Leone schools was also ranked in the first position by the decision tree model. Schools with canteens where the average teachers experience was less than 11.15 years (at Gini impurity of 0.366), were considered in the pass category with 0 Gini impurity.

6.3.2 Private versus government schools

Although government schools employed more experienced and qualified teachers than other schools. Private schools performed better than other schools with an average percentage of papers passed of 69.3 (median: 75.0).

Results showed that 81.8% of private schools (18 out of 22) were located in the western region and only 2 schools were receiving financial support from the government. Schools which were receiving financial support were found to be struggling as only 42.7% passed while 63.2% of non supported schools passed.

Most community schools, government schools and mission schools received financial sup-

port but their average percentage of papers passed was 38.7, 49.9 and 44.8 respectively. Machine learning models associated these types of variables and financial support with a negative impact on performance.

6.3.3 Insignificant variables

Most variables in SL dataset failed to explain the variations in the school outcomes (see Table 5.2 and Figure 5.2). The association of the following school features was not significant with their Goodman-Kruskal's gamma values at P-value > 0.05, namely, remoteness, mixed school, boarding, development plan, drinking water,drinking water source, library, approval status, shift status, garden, internet, private cubicle, science lab, recreational facilities, generator and basic computer skills. Most schools did not have these facilities and there were many missing values for these variables.

Results of ANOVA and Kruskal-Wallis tests showed that the following numeric variables had identical populations for both pass and fail school outcomes thus were not good predictors of school outcomes: total number of latrines, computers, counsellors, chalkboards, textbooks, student-teacher ratio and average teachers experience.

6.4 Summary

This chapter discussed predictive results of trained machine learning models. The performance of models was compared to identity the best algorithm which scored the highest accuracy and Area Under Cover. The best models were selected to extract important features.

Schools were categorised according to their performance in order to investigate the prevalence of both school and community facilities in every category. Frequency tables with percentages were used to explain how these facilities varied across these categories.

Visualizations of feature rankings were also presented to depict model results in easily interpretable way which can be understood by stakeholder. Results showed that determinants of school performance varied across provinces, cities and countries. Schools with improved facilities or located in areas with improved infrastructure performed better than those in disadvantaged environments.

Chapter 7

Discussion

This study focused on discovering determinants of performance in South Africa and Sierra Leone high schools by applying exploratory data analysis tools and machine learning algorithms to school and community level household data.

Chapter 4 explained the properties of both datasets. The South African dataset was mostly community based, while the Sierra Leone dataset mostly contained facilities in the schools. In South Africa, schools were mapped with the most prevalent facilities in their communities. In Sierra Leone, schools were mapped with information about their facilities and exam results, however, the examination dataset contained few schools which reduced the size of the dataset that was used in training machine learning models. High cases of missing data were identified in the Sierra Leone dataset because many schools did not provide the required information in the survey. This problem drastically reduced the number of variables that were considered since instances with missing values could not explain their variations with the performance of schools.

The exploratory data analysis was performed in Chapter 5 to build trust in the datasets as well as discover preliminary findings that provided a baseline for more profound investigation to expand upon. Work in this Chapter 5 guided the process of feature selection where predictor variables were selected to train machine learning models.

Chapter 6 aimed at analysing and interpreting machine learning results. It presents interpretations and visualizations of the model predictive results. Logistic Regression and Tree based algorithms such as: XGBoost, LightGBM, Decision trees and Random Forests were contrasted in performance. The XGBoost model outperformed other machine learning models on both datasets. It adequately explained variations of predictor variables with school performance, hence selected as the best tree based model together with the Logistic Regression models, for purposes of obtaining odds ratios.

This chapter focuses on answering the research questions presented in Chapter 1 using the EDA and predictive results. The discussion focuses on interpreting model results and explaining variations in performance given community or school facilities. Furthermore, brief discussions are presented on the performance of algorithms.

7.1 Insights extracted from South Africa data

There was a significant positive impact of communities features on the education success of schools. Community characteristics were strongly associated with the performance of schools, hence, efforts to improve school or student performance must focus on the community as a whole, not just on the school. Results showed that schools in communities with developed infrastructure and essential facilities such as electricity, water, hospitals in their communities performed better than those in poor communities.

South African schools were categorised according using a quintile system which explained poverty rankings of communities around schools as well as certain infrastructural factors. The poorest quintile is 1 and least poor is 5. The best performing schools belonged to quintile 5 and 4. There was no weak school in the quintile 1 category. Researchers in [51] also found similar performance disparities. Another study in [74] which investigated the impact of community poverty on high school completion for different races found that graduation rates for African American students were more adversely affected by high-poverty environments unlike other student races (white or Asian American students).

Quintiles significantly explained variations with performance of schools in their communities. Moreover, schools in areas with good hospital ratings, access to water and no electricity interruptions performed better than those in communities without these facilities or poor ratings. These results suggest that policymakers should ensure that supportive communities exist even in poor areas to provide schools with a foundation for high academic achievement by expanding the community's capacity to ensure that students' experiences outside school improves the teaching and learning in schools.

7.2 Insights extracted from Sierra Leone data

The analysis and machine learning results linked availability of school canteens and fences, and number of qualified teachers to better school achievements. Private schools performed better than other schools. A further analysis showed that most private schools were fenced, their total number of computers was twice the number of computers in other schools. Private schools also had canteens, electricity grids, libraries and science labs compared to their counterparts. A study in Italy and Spain [77] also had similar results where private schools performed better than public schools. Researchers found that decentralised school funding was associated with higher educational attainments with respect to centralised funding.

7.3 Rural Urban Divide

Schools in urban areas performed better than rural schools in both countries. Analytically, 85.9% of the strong South African schools were located in urban areas compared to 22.0% of fair schools and 12.6% of weak schools. This variable was significant with Goodman-Kruskal's gamma (γ) association coefficient of 0.74 at P-value $< 2.2e^{-16}$ in explaining the variations in school outcomes.

Sierra Leone Schools in the Western region performed better than other regions. This

region was considered urban and it is where the national capital - Freetown is located. Results showed that 83.6% of Western schools had electricity unlike other regions which had less than 45.9% electricity coverage. The number of computers in western region was 3 times more than that in other regions. It was also found that 50.9% of Sierra Leone schools which canteens were in the western region compared to other regions.

Lack of electricity limits the use of electronic teaching materials such as printers which can be used for printing notes, computers and access to internet for research purposes [53]. These facilities support teaching and learning in schools and were found to be determinants of performance in most schools. Researchers in [6] and [25] discussed various factors hindering good performance of rural schools indicating that it is not simply because of their rural locations but because of factors like low teachers' salaries, qualifications and experience.

The performance of schools and quality of education students receive depend on a number of factors such as number of qualified teachers a school can afford. Schools in the western had more qualified teachers. In this region, the average number of teachers with education degrees was 13 and those with any bachelor degree was 22. Comparably, schools in other regions had 8 teachers with education degrees and 10 teachers with any degree. The Southern and North Western regions which were the least performing regions were found to have the least number of qualified teachers with degrees and lowest electricity connectivity. The North Western region was also found to have the least total number of learning materials (textbooks).

7.3.1 Technology, Media and Telecommunications

Technology brings fundamental structural changes that can helpful in achieving significant improvements in performance. It can be used to support teaching and learning by integrating digital learning tools, such as computers, phones and internet.

Considering the Sierra Leone dataset, these school features were not selected for training school performance prediction models due to many missing values. Analytically, they

had low statistical power in explaining school outcomes.

For South Africa, results showed that 84.9% of strong schools were located in communities where most households had DVDs compared to 33.4% of fair schools and 16.8% of weak schools. It was found that more than 88.0% of all schools (strong, fair and weak) were in communities where households had televisions and radios. There was a strong association of access to cell phone internet in communities with school outcomes (γ value of association was 0.63 (P-value $< 2.2e^{-16}$). After examining the underlying dataset, results showed that 46.8% of the strong schools were in communities that had access to internet compared to 14.6% and 8.4% of fair and weak schools respectively. Work in [24, 26, 69] also associates these services to better performance and improved access to information which in this case maybe relevant to school candidates.

7.3.2 Security

Availability of school fence in Sierra Leone schools was associated with more cases of passing as 55% of fenced schools passed compared to the 38% unfenced schools. A school fence is established to prevent or control access or exits can be considered as security measure with the aim of not only protecting the students and staff, but also reducing events of students escaping from schools and student misbehavior [68].

Like wise, security in SA communities was linked to school outcomes as low cases of very-safe safety during day could have a negative impact on performance. It was further noticed that over 91% of the struggling schools were located in very safe communities during day compared to only 75.5% of strong schools.

Rural areas were found to be safer than urban areas during day and night. Security results showed that 96.5% of schools in rural areas were in very safe places compared to 67.3% of urban schools. The LR model indicated that very-safe safety during day was associated to increasing the odds of passing over failing by 3.43, while communities with unsafe nights were linked to reducing the odds of passing by 0.47 (-53.36%), when other factors remain constant.

7.4 Policy Implication

This section reflects on the connections between resources in schools and communities with education achievements. These insights can support enforcement of suitable and viable education policies which seek to improve learning results.

Although most features were found to be linked to the performance in schools, there was no strong connections between some facilities and school performance. Weak relationships makes it difficult to determine an objective strategy for deciding what is sufficient to improve schools. For instance, in Sierra Leone, there were more experienced and qualified teachers in government schools than other schools, but private schools performed better than government schools.

Teaching experience is positively associated with student achievement gains throughout in schools [1]. Most experienced instructors provide extensive support to student learning and to fellow teachers in schools, but this was not enough for government schools to perform better. Various conceivable reasons such as inefficient operations, poor leadership and lack of enough information about existing or missing facilities in schools can lead poor resources management and performance.

Effective allocation of resources based on well-defined structures poses a greater likelihood of success, but there is no clear definition of what is adequate to support teaching and learning because it is seen as a political or financial issue which changes with different political views and national economic demands [29]. Improving school facilities as well as increasing scholastic supplies is not enough to provide a noticeable stride in performance but can be effective if it is enhanced by effective management in the school. For instance, only 42.7% of schools which received financial support managed to help their students pass more than 50% of the papers, while 63.2% of financially non-supported schools passed the performance threshold.

Results showed that most private schools which did not receive financial support performed better than government and mission schools which received more support. Monitoring which facilities most finances are spent on is more important than exaggerating

the amount spent especially in schools experiencing poor leadership and corruption. Most private schools had canteens, electricity grids, libraries, science labs and school fences. Their total number of computers was twice the number of computers in other schools. Some of this factors were found to be linked to better performance hence worthy spending school money on.

In order to enact appropriate education policies, policy makers should first seek to understand the circumstances under which schools operate and how they utilise their resources and finances. This step provides relevant information about underlying policies and practices guiding how and which resources are provided to attain desired teaching and learning outcomes. For instance, providing unnecessary equipment and facilities with schools may not be sufficient to improve schools and performance if they are not matched to address particular learning needs, and if there's no capacity built to successfully oversee those resources. Schools should be entrusted to uphold programs and processes which ensure expertise in the administration of resources by encouraging by encouraging financial audits, monitoring and supervision.

7.5 Recommendations for data collection

Big data educational research requires enough interlinked educational datasets to support policy making and improve schools. Education stakeholders should encourage processes which promote the collection of various layers of information about school systems, students and their surrounding environments. Examination data and information about facilities in school with their surrounding communities should be made freely accessible to everybody through reports with straightforward details of how these facilities are utilized with given prove of their effect on learning results. Data sources must provide all relevant information to avoid missing data points. Data should contain socioeconomic backgrounds of students and teachers, teaching and learning processes, school leadership and details of access to or usage of school amenities. This information can be used to identify areas which require urgent attention and improve schools.

7.6 Summary

This chapter discussed the insights that answer the research questions posed in Chapter 1.

The insights drawn from data analysis and interpretation of machine learning model results were addressed to explain their association with school performance. Similar results from other studies were contrasted with the findings of this study.

A discussion was presented on the policy implications of the results of this study and what policymakers should consider when improving resource allocation schools.

Lastly, recommendations for data collection were discussed highlighting important data aspects which should be collected and shared.

Chapter 8

Conclusion

Big educational datasets offers unprecedented opportunities for schools and policy-makers to understand how students learn from different educational events and what takes place in and around school systems. This research aimed to identify determinants of school performance in South Africa and Sierra Leone using data mining and machine learning techniques presented in Chapter 3. Based on the results, it can be concluded that improving resource allocation in schools and developing communities around schools can potentially improve the performance of students. Results indicate that school locations, availability of canteen, fence, electricity, and qualified teachers in schools were significantly associated with the performance of schools.

Multiple existing datasets were merged to form a complete dataset for each country. These included community household survey datasets, school census data, examination results and teachers' datasets. The Sierra Leone dataset was rich with information about schools and their facilities. South Africa's school master lists were similar to Sierra Leone school census data, but lacked enough information about schools facilities. However, for South Africa, community household survey data was used to investigate the performance of schools located in these communities. All datasets were not subjected to an ethical clearance process and were acquired without sensitive or identifying information from government agencies. Chapter 4 explains the properties of these datasets, while

Chapter 5 discussed high level exploratory data analysis results which helped to summarise distributions and relationships within the data as well as raising confidence in the datasets.

In attempt to improve schools, results of this study were purposely to guide resource allocation and inform educational policy making. Like other statistical analysis findings, this study discovered similar factors associated with school performance. Determinants of performance varied across countries as presented in Chapter 6 and discussed in Chapter 7. These performance determinants can be used to support the fact that improving the school and community environment can potentially influence the performance of students in schools.

While results show that schools which were located in developed communities performed better and had access to more facilities, this research provided insights into the management of the resources in or around schools. Building capacities to effectively manage resources in schools and their communities is critical to ensure that they particularly serve individual needs required to improve performance.

8.1 Summary Of Conclusions

This section provides an account of highlight findings from this study, a more detailed discussion is provided in Chapter 7.

- The XGBoost algorithm outperformed other algorithms with high accuracy and Area Under Curve. This performance was expected because XGBoost uses a gradient boosting technique unlike Decision trees and Random forest algorithms.
- Schools in urban areas performed better than rural schools in both countries. Rural schools did not perform poorly because of their location, but they were found to be located in areas which lacked enough basic facilities.
- The performance of schools in South Africa was highly determined the quintile of

the community where the school is located. The top best performing schools were located in quintile 5 communities, while poor performing schools were located in quintile 1 communities. Quintile 5 communities are regarded rich and have developed infrastructure compared to quintile 1 communities.

8.2 Future Work

Based on conclusions of this study, similar studies should be conducted in other countries to customise educational policies based on the complexities and differences in resource variations with performance in their schools.

Results in this study should be viewed along side other social scientists and educationists work which explain social issues in education. For instance, how quintile factors, rural-urban divides, and safety issues affect performance.

More educational datasets should be collected and publicly made available to support educational research.

Lastly, more useful data-driven exploratory tools powered by machine learning interpretable models should be developed to assist in viewing these relationships, quantifying resource inputs and school outcomes in a predictive approach for implementation with policy makers who seek to transform education outcomes at a national and local level.

Bibliography

[1] TO Adeyemi. Teachers' teaching experience and students' learning outcomes in secondary schools in ondo state, nigeria. *Asian journal of information technology*, 7(5):201–209, 2008.

[2] Sunday A Adeyemo. The relationship among school environment, student approaches to learning and their academic achievement in senior secondary school physics. *International journal of educational research and technology*, 3(1):21–26, 2012.

[3] Mostafa Al-Emran and Said A Salloum. Students' attitudes towards the use of mobile technologies in e-evaluation. *International Journal of Interactive Mobile Technologies (IJIM)*, 11(5):195–202, 2017.

[4] Hanan Aldowah, Hosam Al-Samarraie, and Wan Mohamad Fauzy. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37:13–49, 2019.

[5] Kate Algozzine and Bob Algozzine. Classroom instructional ecology and school-wide positive behavior support. *Journal of Applied School Psychology*, 24(1):29–47, 2007.

[6] Funmilola Bosede Alokan and Amos Emiloju Arijesuyo. Rural and urban differential in student's academic performance among secondary school students in ondo state, nigeria. *Journal of Educational and Social Research*, 3(3):213, 2013.

[7] Mohamed Aly. Survey on multiclass classification methods. *Neural Netw*, 19:1–9, 2005.

[8] AS Arul Laurence. School environment & academic performance of standard six students. *Journal of Educational and Industrial Studies in the World*, 2(3):22–27, 2012.

[9] Ryan SJD Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining*, 1(1):3–17, 2009.

[10] David Blazar and Matthew A Kraft. Teacher and teaching effects on students' attitudes and behaviors. *Educational evaluation and policy analysis*, 39(1):146–170, 2017.

[11] Stephanie L Blumenshine, William F Vann Jr, Ziya Gizlice, and Jessica Y Lee. Children's school performance: impact of general and oral health. *Journal of public health dentistry*, 68(2):82–87, 2008.

[12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[14] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan. Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, 7(3):246–259, 2014.

[15] Kathleen Cotton and Karen Reed Wikelund. Schoolwide and classroom discipline. *School Improvement Research Series*, 9:1–28, 1990.

[16] Ben Daniel. B ig d ata and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, 46(5):904–920, 2015.

[17] James A Davis. A partial coefficient for goodman and kruskal's gamma. *Journal of the American Statistical Association*, 62(317):189–193, 1967.

[18] Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[19] Veerle Dieltiens and Shireen Motala. Quintile ranking system, south africa. *Achieving transparency in pro-poor education incentives*, page 69, 2014.

[20] Martin Dougiamas and Peter Taylor. Moodle: Using learning communities to create an open source course management system. In *EdMedia+ Innovate Learning*, pages 171–178. Association for the Advancement of Computing in Education (AACE), 2003.

[21] S Eric. The role of supportive school environment in promoting success, an article from development studies centre (dsc). *Developing Safe and Healthy Kids, Published in Getting Result*, 2005.

[22] Rebecca Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6):304–317, 2012.

[23] Mahesh Gadhavi and Chirag Patel. Student final grade prediction based on linear regression. *Indian J. Comput. Sci. Eng.*, 8(3):274–279, 2017.

[24] Kiran Gaikwad, Gaurav Paruthi, and William Thies. Interactive dvds as a platform for education. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, pages 1–10, 2010.

[25] Robert Gibbs. The challenge ahead for rural schools. In *Forum for Applied Research and Public Policy*, volume 15, page 82. University of Tennessee, Energy, Environment and Resources Center, 2000.

[26] Susan Gibson and Dianne Oberg. Visions and realities of internet use in schools: Canadian perspectives. *British Journal of Educational Technology*, 35(5):569–585, 2004.

[27] Alaa Hamoud, Ali Salah Hashim, and Wid Akeel Awadh. Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5:26–31, 2018.

[28] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[29] Eric A Hanushek. Assessing the effects of school resources on student performance: An update. *Educational evaluation and policy analysis*, 19(2):141–164, 1997.

[30] Hind Bushra Ahmed Ibrahim. The role of school feeding program supported by dal company in students' enrolment and drop-out. *Advances in Social Sciences Research Journal*, 4(2), 2017.

[31] Gareth James and Trevor Hastie. The error coding method and picts. *Journal of Computational and Graphical statistics*, 7(3):377–387, 1998.

[32] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[33] Stephanie J Jones. Technology review: the possibilities of learning analytics to improve learner-centered decision-making. *Community College Enterprise*, 18(1):89–93, 2012.

[34] Reynol Junco and Candrianna Clem. Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, 27:54–63, 2015.

[35] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[36] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[37] Matthew A Kraft, William H Marinell, and Darrick Shen-Wei Yee. School organizational contexts, teacher turnover, and student achievement: Evidence from panel data. *American Educational Research Journal*, 53(5):1411–1449, 2016.

[38] Matthew J Kruger-Ross and Richard D Waters. Predicting online learning success: Applying the situational theory of publics to the virtual classroom. *Computers & Education*, 61:176–184, 2013.

[39] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[40] Dileep Kumar. Tree-based modeling techniques. In *Machine Learning Techniques for Improved Business Analytics*, pages 1–18. IGI Global, 2019.

[41] S Anupama Kumar et al. Efficiency of decision trees in predicting student's academic performance. 2011.

[42] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[43] Jindřich Matoušek and Daniel Tihelka. Using extreme gradient boosting to detect glottal closure instants in speech signal. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6515–6519. IEEE, 2019.

[44] Donald Michie, David J Spiegelhalter, CC Taylor, et al. Machine learning. *Neural and Statistical Classification*, 13(1994):1–298, 1994.

[45] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

[46] Lebusa Monyooe. Reclassifying township schools—south africa's educational tinkering expedition! *Creative Education*, 8(03):471, 2017.

[47] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[48] Angela Nebot, Félix Castro, A Vellido, and Francisco 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*, pages 74–79, 2006.

[49] Edward C Norton, Bryan E Dowd, and Matthew L Maciejewski. Odds ratios—current best practice and use. *Jama*, 320(1):84–85, 2018.

[50] Ministry of Education Science and Technology. Education sector plan 2018-2020. sierra leone, 2018. Accessed on: 2018-11-28.

[51] Ugorji I Ogbonnaya and Francis K Awuah. Quintile ranking of schools in south africa and learners' achievement in probability. 2019.

[52] Haidar Osman, Mohammad Ghafari, and Oscar Nierstrasz. Hyperparameter optimization to improve bug prediction accuracy. In *2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE)*, pages 33–38. IEEE, 2017.

[53] Mohamed Ouf, Mohamed H Issa, Phil Merkel, and Panos Polyzois. The effect of occupancy on electricity use in three canadian schools. *Journal of Green Building*, 13(1):95–112, 2018.

[54] Joseph Sunday Owoeye and Philias Olatunde Yara. School location and academic achievement of secondary school in ekiti state, nigeria. *Asian social science*, 7(5):170–175, 2011.

[55] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[56] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[57] Simone Robers, Jijun Zhang, and Jennifer Truman. Indicators of school crime and safety: 2011. nces 2012-002/ncj 236021. *National center for education statistics*, 2012.

[58] Amanda J Rockinson-Szapkiw, Jennifer Courduff, Kimberly Carter, and David Bennett. Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education*, 63:259–266, 2013.

[59] Cristobal Romero, Pedro G Espejo, Amelia Zafra, Jose Raul Romero, and Sebastian Ventura. Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146, 2013.

[60] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472, 2013.

[61] Cristobal Romero and Sebastian Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[62] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[63] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[64] Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. *Advances in neural information processing systems*, 16:1237–1244, 2003.

[65] Victoria Rubin and Tatiana Lukoianova. Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24(1):4, 2013.

[66] Barbara Schneider and Yongsook Lee. A model for academic success: The school and home environment of east asian students. *Anthropology & Education Quarterly*, 21(4):358–377, 1990.

[67] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[68] Timothy J Servoss. School security and student misbehavior: A multi-level examination. *Youth & Society*, 49(6):755–778, 2017.

[69] Iman Sharif and James D Sargent. Association between television, movie, and video game exposure and school performance. *Pediatrics*, 118(4):e1061–e1070, 2006.

[70] Ronald E Shiffler. Maximum z scores and outliers. *The American Statistician*, 42(1):79–80, 1988.

[71] George Siemens. The journal of learning analytics: Supporting and promoting learning analytics research. *Journal of Learning Analytics*, 1(1):3–5, 2014.

[72] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[73] Matthew P Steinberg, Elaine Allensworth, and David W Johnson. *Student and Teacher Safety in Chicago Public Schools: The Roles of Community Context and School Social Organization.* ERIC, 2011.

[74] Christopher B Swanson et al. Who graduates? who doesn't?: A statistical portrait of public high school graduation, class of 2001. 2004.

[75] Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227, 2010.

[76] Martha Tapia, George E Marsh, et al. An instrument to measure mathematics attitudes. *Academic Exchange Quarterly*, 8(2):16–22, 2004.

[77] Gilberto Turati, Daniel Montolio, and Massimiliano Piacenza. Funding and school accountability: The importance of private and decentralised public funding for pupil attainment. *Fiscal Studies*, 38(4):615–639, 2017.

[78] Erlanger A Turner, Megan Chandler, and Robert W Heffer. The influence of parenting styles, achievement motivation, and self-efficacy on academic performance in college students. *Journal of college student development*, 50(3):337–346, 2009.

[79] Shahadat Uddin, Kate Thompson, Beat Schwendimann, and Mahendra Piraveenan. The impact of study load on the dynamics of longitudinal email communications among students. *Computers & Education*, 72:209–219, 2014.

[80] Mudassir Ibrahim Usaini, Norsuhaily Binti Abubakar, and Ado Abdu Bichi. Influence of school environment on academic performance of secondary school students in kuala terengganu, malaysia. *The American Journal of Innovative Research and Applied Sciences*, 1(6):203–209, 2015.

[81] Chris Van Wyk. An overview of key data sets in education in south africa. *South African Journal of Childhood Education*, 5(2):146–170, 2015.

[82] Gottfried Vossen. Big data as the new enabler in business and other intelligence. *Vietnam Journal of Computer Science*, 1(1):3–14, 2014.

[83] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Springer-Verlag London, UK, 2000.

[84] Yu Xiaogao and Peng Ruiqing. Research on big data-driven high-risk students prediction. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 145–149. IEEE, 2017.

[85] Ibrahim Yildirim. The effects of gamification-based teaching practices on student achievement and students' attitudes toward lessons. *The Internet and Higher Education*, 33:86–92, 2017.