Modelling Student Retention in Tutorial Classes with Uncertainty – A Bayesian Approach to Predicting Attendance-based Retention

Eli Bila Nimy

Supervisor(s):
Dr. Moeketsi Mosia

A mini-dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

Department of Computer Science and Information Technology
Sol Plaatje University, Kimberley

19 November 2023

Declaration

I, Eli Bila Nimy, declare that this mini-dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at Sol Plaatje University, Kimberley. It has not been submitted for any degree or examination at any other university.

Eli Bila Nimy
19 November 2023

Abstract

Bayesian additive regression tree (BART) is a recent statistical method that blends ensemble learning with nonparametric regression. BART is constructed using a Bayesian approach, which provides the benefit of model-based prediction uncertainty, enhancing the reliability of predictions. This study proposes the development of a BART model with a binomial likelihood to predict the percentage of students retained in tutorial classes using attendance data. The proposed model is evaluated and benchmarked against the Random Forest Regressor (RFR). The proposed BART model reported an average of 20% higher predictive performance compared to RFR across five error metrics, achieving an R-squared score of 0.9414. Furthermore, the study demonstrates the utility of the Highest Density Interval provided by the BART model, which can help in determining the best and worst-case scenarios for student retention rate estimates. The significance of this study extends to multiple stakeholders within the educational sector. Educational institutions, administrators, and policymakers can benefit from this study by gaining insights into how future tutorship programme student retention rates can be predicted using predictive models. Moreover, the foresight provided by the predicted student retention rates can aid in strategic resource allocation, facilitating more informed planning and budgeting for tutorship programmes.

Keywords: Student retention, Tutorship programme, Attendance, Educational data analytics, Bayesian, Regression, Ensemble

Acknowledgements

I would like to thank my research supervisor Dr. Moeketsi Mosia, for his patience and assistance throughout the planning and writing phases of the research proposal and report, even in a time of high uncertainty, his support did not waver. My thanks also go to my family, friends, and class fellows for the honest peer reviews, feedback, and support.

The support of the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research report is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NEPTTP.

Contents

List of Figures

List of Tables

List of Abbreviations

TP	**T**utorship **P**rogramme
SPU	**S**ol **P**laatje **U**niversity
CTLPD	**C**entre for **T**eaching, **L**earning, and **P**rogramme **D**evelopment
KDD	**K**nowledge **D**iscovery in **D**atabase
AA	**A**cademic **A**nalytics
EDM	**E**ducational **D**ata in **M**ining
LA	**L**earning **A**nalytics
MLE	**M**aximum **L**ikelihood **E**stimation
MAP	**M**aximum **A** **P**osterior
BART	**B**ayesian **A**dditive **R**egression **T**ree
RFR	**R**andom **F**orest **R**egressor
HDI	**H**igest **D**ensity **I**nterval
MAE	**M**ean **A**bsolute **E**rror
MedAE	**Med**ian **A**bsolute **E**rror
RMSE	**R**oot **M**ean **S**quared **E**rror
CP	**C**onformal **P**rediction

Chapter 1

Introduction

In today's higher education landscape, retaining students has emerged as a significant challenge. Statistics reveal that nearly 40% of students are not retained within their academic institutions or classes over time, indicating the urgency of addressing this issue [32]. Consequently, educators and researchers have undertaken numerous studies to explore innovative approaches to student retention. One approach that has gained significant attention is data analytics. Educational institutions have begun recognizing the potential of data analytic solutions in ensuring student success and retention [36]. Student retention, in this context, is defined as the percentage of students who re-enroll from one academic year to the next [32]. The term data analytics refers to the systematic examination of raw data within a specific context to uncover meaningful patterns, correlations, and trends that can be translated into actionable insights [26]. The application of data analytics in the education domain is referred to as educational data analytics [26].

Recent studies demonstrate that higher education institutions that develop predictive and diagnostic analytical solutions to address student retention can benefit from enhanced reputation, better ranking, and financial stability [9, 26, 35]. As low retention rates negatively impact an institution's financial and institutional stability [9, 28], exhibiting the need for educational institutions to develop and implement effective data analytic strategies to improve student retention rates.

While existing work in educational data analytics has primarily focused on the development of predictive models, such as logistic regression, support vector machines, random forest, and decision trees, for the purpose of predicting student retention, a notable gap exists in the consideration of student retention as a multifaceted problem [3, 9, 32]. Student retention encompasses more than just predicting whether a student can be retained from one year to the next; it also extends to various programmes offered by institutions, including tutorship programmes (TPs).

TPs have become an essential part of higher educational institutions as they provide a supportive environment for students to improve their academic performance [38]. TPs offer personalized attention, access to resources, and mentorship opportunities, which can enhance students' understanding of complex concepts and help develop their study skills [38]. Moreover, TPs can increase students' confidence in their abilities, leading to improved academic and personal outcomes [38]. However, to ensure the effectiveness of tutorship programmes, it is crucial to ensure retention in tutorial classes. This study defines student retention in tutorial classes as the percentage of students retained in tutorial classes over a specified period. High retention rates in TPs not only ensure that students continue to benefit from personalized instruction and support but also help to improve graduation rates [7]. In this way, TPs play a critical role in improving student success and supporting their academic and personal development [7].

Despite the progress made in using educational data analytics to tackle student retention, gaps in literature persist. One significant gap pertains to the absence of predictive models in education that can offer reliable predictions using uncertainty quantification approaches [5]. The current predictive models used in studies lack probabilistic information, which leads to overly confident or incomplete decision-making. Uncertainty quantification provides a framework for estimating and integrating probabilistic information into predictive models [27]. By disregarding uncertainty, these studies fail to assess the confidence associated with predictions. Uncertainty information is crucial for decision-making and risk assessment. Ignoring uncertainty can result in incomplete or deceptive decision-making [24]. Decision-makers in education need not only point predictions but also a measure of certainty in those predictions. Neglecting uncertainty may result in underestimating risks, overestimating benefits, or making suboptimal choices [24].

To address these gaps, this study proposes the adoption of a Bayesian approach to model and predict attendance-based retention in tutorial classes. Bayesian models, grounded in probability theory, provide a statistical framework for making predictions and decisions [18]. Notably, Bayesian models offer several benefits over traditional predictive models such as logistic regression, support vector machines, random forest, and decision trees. These benefits include the ability to incorporate prior information, quantify uncertainty and handle limited data [24].

1.1 Problem Statement

The study is designed to bridge a theoretical gap and solve a practical problem in the education data analytics domain. Firstly, it aims to fill the scholarly void surrounding the application of a Bayesian approach with uncertainty quantification in the education domain, which has received limited attention thus far [5]. This prompts the need to empirically examine how a Bayesian approach can reliably predict and offer actionable insights into the student retention problem.

Secondly, the study seeks to contribute towards resolving a practical problem faced by the Centre for Teaching, Learning, and Programme Development (CTLPD) at Sol Plaatje University (SPU). Currently, the CTLPD lacks an effective decision-making and resource allocation model for its tutorship programme, resulting in inefficiencies and potential costs. To address this issue, this study explores the implementation of a Bayesian model that will enable the CTLPD to reliably predict future retention rates in tutorial classes and proactively identify opportunities for intervention to improve retention rates.

1.2 Research Questions

- RQ1: To what extent does the performance of a Bayesian model differ from that of a non-Bayesian model in terms of predicting retention in tutorial classes, as evaluated using metrics including minimum error, maximum error, mean absolute error, median absolute error, root mean squared error, and coefficient of determination?
- RQ2: How can the highest density interval be used to summarize the uncertainty in retention predictions and make inferences about retention in tutorial classes?

1.3 Research Aims and Objectives

1.3.1 Research Aim

To develop a Bayesian model to predict attendance-based retention in tutorial classes.

1.3.2 Objectives

In line with the aim of this study, the following objectives drive this study:

- Develop a Bayesian model using attendance data from tutorship programmes to predict retention in tutorial classes.
- Compare the performance of a Bayesian and non-Bayesian model in predicting retention in tutorial classes based on the minimum error, maximum error, mean absolute error, median absolute error, root mean squared error, and coefficient of determination.
- Use the highest density interval to summarize the uncertainty in retention predictions and draw inferences about retention in tutorial classes.

1.4 Scope and Limitations

The primary focus of this study is to offer a fresh perspective on the student retention problem by highlighting its potential existence within student support programmes. These programmes are designed as interventions to enhance student success, with retention being a pivotal factor for students to derive maximum benefit from such initiatives. However, it is imperative to acknowledge a noteworthy limitation in the scope of this work, as it exclusively examines student retention through the lens of a singular student support programme, namely a tutorship programme. The data used for data analysis and modeling is derived solely from tutorial class attendance data at Sol Plaatje University.

Despite this limitation, the methodology employed in this study underscores the importance of two key variables – student numbers for attendance tracking and programme attendance dates. These variables are identified as critical components for implementing the proposed predictive models for student retention. Importantly, any educational institution equipped with the capability to collect these two essential variables stands to benefit substantially from the methodology outlined in this study.

While the specific focus on a tutorship programme within a single institution imposes a limitation on the study's external validity, the identified crucial variables make the proposed methods applicable and beneficial to educational institutions that have the capacity to collect comparable attendance data.

1.5 Layout of Chapters

- Chapter 2 – The literature review chapter comprehensively delves into the three primary themes underpinning this study: educational data analytics, Bayesian modelling, and predictive analytics for modelling student retention. The various types of data analytics within the education domain are explained, supported by examples, highlighting the types employed in this study. This chapter further explains the Bayesian modelling approach, offering insights into the construction of Bayesian models, Bayesian inference, and the methods used by non-Bayesian models for parameter estimation.

- Chapter 3 – The methodology chapter presents a detailed account of the research method used, offering a step-by-step walkthrough of the Knowledge Discovery in Database (KDD) framework, which played an instrumental role in achieving the study's objectives. The KDD framework encompasses four pivotal steps: (1) data collection and understanding, (2) data preprocessing and transformation, (3) modelling, and (4) evaluation.

- Chapter 4 – The results and discussion chapter is organised into three sections. First, data analysis results on the tutorial class attendance data are presented and discussed. Subsequently, the model performance results are presented and discussed. Finally, the application of the highest density interval of the Bayesian model to quantify uncertainties in student retention predictions is detailed.

- Chapter 5 – The conclusion chapter is split into two key sections. The first section addresses the research questions, offering conclusive insights. The second section directs attention to future research, outlining ways in which this study can be extended.

Chapter 2

Literature Review

This chapter reviews the relevant literature for the three main topics of this study. Firstly, the types of data analytics in the education domain are discussed, highlighting the type(s) adopted in this study. Secondly, the Bayesian modeling approach is explained. Finally, existing work on the use of predictive analytics for modeling student retention is presented, with a focus on demonstrating the commonly used models and variables.

2.1 Types of Data Analytics in Higher Education

The expansion of data analytics in higher education is being driven by the necessity to create innovative solutions based on data to address the challenges faced in education [27, 26]. This trend is further fueled by the growing amount and diversity of data collected from both online and traditional university offerings, opening up new possibilities for using data analytics to enhance the quality of higher education [36]. Consequently, different terms that are closely related, such as Academic Analytics (AA), Educational Data Mining (EDM), and Learning Analytics (LA), have emerged to represent distinct types of data analytics employed in higher education [26]. These terms indicate various approaches to data analytics used in the field. Furthermore, the outcomes of one type of data analytics can serve as input for another, resulting in a complex and interconnected landscape of data analytics approaches in higher education. This section will explore each type of data analytics in education and determine which ones will be used in this study.

2.1.1 Academic Analytics

Academic Analytics is a term that is defined as "the application of data analytic techniques and tools for purposes of supporting institutional operations and decision-making" [26, p. 67]. The primary focus of Academic Analytics is to enhance institutional operations and decision-making processes. This process involves the use of data analytic techniques and tools at five distinct levels, namely faculty, institutional, regional, national, and international levels [26]. It is worthy to note that Academic Analytics offers potential benefits to a diverse range of individuals and groups, including students, faculty, and executive officers.

The utilization of Academic Analytics can bring significant advantages to faculty members. Through the examination of educational data, Academic Analytics has the capability to provide important factors that contribute to student success, offer valuable insights into effective methods, and enhance knowledge about teaching and learning [8, 34]. Student success holds a prominent position as a key performance indicator (KPI) in higher education, and therefore, most faculty members are highly interested in predicting and monitoring student success. Studies have shown that student engagement indicators, such as attendance, clicks, and time spent on learning management systems (LMS), are crucial predictors of student success [27]. Using Academic Analytics, faculty members can gain access to this information and use it to inform their teaching practices.

2.1.2 Educational Data Mining

Education Data Mining is defined as "the development and evaluation of data analytics methods for exploring educational data and using those methods to better understand learners and the learning environment" [26, p. 67]. The primary objects of interest within the field of EDM are the methods and techniques employed for the purpose of analyzing data at various levels within the educational system, namely departmental, faculty, and institutional levels [26]. The various methods and techniques applied in EDM have been categorized in five general groups. These groups are clustering, relationship mining, prediction, discovery with models, and distillation of data for human judgement [4, 22]. Prediction methods are used to forecast future outcomes, while clustering methods are applied to identify groups with similar attributes [26]. Relationship mining explores correlations between different variables, and discovery with models aims to uncover hidden patterns in the data [26]. The final group, distillation of data for human judgement, involves summarizing complex data into easily interpretable formats that can aid in decision-making [26].

2.1.3 Learning Analytics

Learning Analytics refers to "the application of data analytic techniques and tools for purposes of understanding and enhancing learning and teaching" [26, p. 67]. The primary focus of learning analytics is the learners and the learning settings, which are subject to data analysis at the levels of individual students, courses, and departments [26].

As per the Society for Learning Analytics Research (SOLAR), learning analytics can be categorized into four distinct areas, which include descriptive, diagnostic, predictive, and prescriptive analytics [33]. Descriptive analytics provides insights into past events, and it can be achieved through the examination of student feedback from surveys, as well as data that describes the student's lifecycle, such as study support, enrollments, and exams [33]. Diagnostic analytics, on the other hand, aims to identify underlying patterns in the data. This type of analytics is achieved by analyzing educational data to find key performance indicators and metrics that can be used to enhance student engagement [33]. Predictive analytics focuses on understanding the future by identifying patterns in historical data and utilizing statistical models and algorithms to capture relationships and forecast future outcomes. Examples of predictive analytics in learning analytics include predicting at-risk students, student drop-out rates, and retention rates [32]. Lastly, prescriptive analytics aims to offer advice on potential outcomes and recommend choices using machine learning and business rules [33]. Through this type of analytics, institutions can make informed decisions on the best course of action to take, given the available data.

The various forms of data analytics in higher education vary in terms of their focus and the level of the education system they target. It has been noted before that the results obtained from one type of educational data analytics can be used as input for another. In this study, the primary approach to data analytics employed is a combination of academic analytics, which focuses on institutional operations and decision making at an institutional level, and educational data mining, which involves predicting student retention in tutorial classes also at an institutional level.

2.2 Bayesian Modeling

Bayesian modelling, grounded in the principles of probability theory, provides a sophisticated and principled approach to dealing with uncertainty and incomplete information [19, 23]. This section delves into the concept of modelling and the underpinnings of Bayesian modelling and its distinction from non-Bayesian models. The non-Bayesian subsection provides a clear distinction on the model estimation methods used in Bayesian and non-Bayesian models. Mathematical formulas are used to elucidate these concepts.

2.2.1 Bayesian Models

In the space of research and practice, models are simplified descriptions of a system or process. Models are designed to deliberately encompass the most significant or relevant variables of a system [18].

Computationally or otherwise Bayesian models have two defining characteristics [19]:

- Probability distributions: Probability distributions are used to represent unknown quantities, known as parameters.
- Bayes theorem: Bayes theorem is employed as a mechanism to update the parameter values based on the available data.

At a high level, constructing Bayesian models involves three main steps [19]:

- Creating a model by combining and transforming random variables, based on assumptions about how the data was generated, using available data.

- Using Bayes theorem to condition the model to the available data. This process is called inference, resulting in the posterior distribution. While this step is expected to reduce uncertainty in possible parameter values, it is not guaranteed.
- Critiquing the model by evaluating whether it aligns with different criteria, such as the available data and domain-knowledge expertise. This step is necessary due to the uncertainties that practitioners or researchers may have about the model, sometimes requiring comparison with other models.

2.2.2 Bayesian Inference

Put simply, inference involves drawing conclusions using evidence and reasoning [23]. Bayesian inference is a particular form of statistical inference where probability distributions are combined to derive updated distributions [23]. The process relies on Bayes theorem to estimate the value of a parameter θ based on observed data Y.

$$p(\theta|Y) = \frac{p(Y|\theta) \cdot p(\theta)}{p(Y)} \tag{2.1}$$

The concept of likelihood $p(Y|\theta)$ involves incorporating data into the model, while the prior distribution $p(\theta)$ represents knowledge about the parameters θ prior to observing the data Y. The posterior distribution $p(\theta|Y)$, which combines the likelihood and prior distribution, captures all the relevant information about the problem. The marginal likelihood $p(Y)$, which represents the probability of observing the data across all possible parameter values, is often not computed. As a result, Bayes theorem is typically expressed as a proportionality [19]:

$$p(\theta|Y) \propto p(Y|\theta) \cdot p(\theta) \tag{2.2}$$

In Bayesian inference, a useful quantity to compute is the posterior predictive distribution [19]:

$$p(\hat{Y} \mid Y) = \int p(\hat{Y} \mid Y) \cdot p(\theta \mid Y)\, d\theta \tag{2.3}$$

The posterior predictive distribution refers to the distribution of future data, $\hat{Y}$, that is expected based on the posterior $p(\theta \mid Y)$, which is derived from the model (comprised of the prior and likelihood) and observed data. Essentially, this represents the data that the model predicts will be seen after analyzing the dataset. The equation for the posterior predictive distribution involves integrating over the posterior distribution of parameters, which means that predictions are made while taking into account the uncertainty associated with model estimates.

2.2.3 Non-Bayesian Models

In contrast to Bayesian models, non-Bayesian models typically estimate model parameters using frequentist techniques such as maximum a posteriori estimation (MAP) or maximum likelihood estimation (MLE) [23]. These methods involve identifying a single-point estimate that maximizes the likelihood of observing the data or the posterior probability respectively [25, 23]. Unlike Bayesian models, frequentist models do not yield a full probability distribution over the model parameters, which can hinder the ability to quantify and propagate uncertainty.

The point estimate θ in MLE is derived by maximizing the likelihood:

$$\theta = \arg\max p(D \mid \theta) \tag{2.4}$$

The point estimate θ in MAP is derived by maximizing the posterior probability:

$$\theta = \arg\max p(\theta \mid D) \tag{2.5}$$

Bayesian modeling, on the other hand, seeks to compute the full posterior distribution over the model parameters, given the observed data [23]. This approach enables a more comprehensive representation of uncertainty and allows for principled decision-making under uncertainty [19]. Furthermore, Bayesian models can naturally incorporate prior information, which can be particularly useful when dealing with limited or noisy data [27].

2.3 The Use of Predictive Analytics in Modeling Student Retention

This section focuses on the application of predictive analytics in enhancing student retention based on previous research. It discusses the predictive models employed and important factors considered when modeling student retention within the context of higher education.

The potential of data mining methods for developing predictive models to manage student retention in higher education was proposed by Yadav, Bharadwaj, and Pal [39]. The primary aim was to identify students who require help from the student retention programme. The researchers implemented three decision tree classification models: ID3, C4.5, and ADT. Their findings indicated that the inclusion of all social, personal, environmental, and psychological variables is vital for effective prediction of student retention rates. The variables used in the models included gender, student category, secondary school grades, secondary school math grade, graduation stream, graduation grade, medium of teaching, college location, admission type, and retention.

The effectiveness of predictive deep learning techniques in analyzing student learning data and predicting student retention was demonstrated by Uliyan et al. [36]. The researchers utilized the bidirectional long short-term model (BLSTM) and condition random field (CRF) deep learning techniques, which accurately predicted student retention. The researchers benchmarked these deep learning techniques against several other models, including neural network, decision tree, random forest, naïve bayes, support vector machines, and logistic regression. Evaluation metrics such as recall, accuracy, precision, and F-score were employed to assess the models' performance. The predictive variables used to forecast retention included preparatory grade-point (GPA), mathematics, physics, English, quizzes, assignments, statistics grade, high school, and overall GPA. The study concluded that predictive models can be valuable tools for universities to determine students at risk of discontinuing their studies.

The use of support vector machines and neural networks models to predict student retention was explored by Trivedi [35] with impacts and implications. The study used degree, gender, age, 1st generation, high school GPA, college GPA, plans to work, and ACT composite as input variables for the models. Interestingly, the authors found that high school rank, first math course grade, SAT math score, and pre-college intervention programmes were useful in predicting retention. This suggests that non-academic factors, such as preparation programmes, may have an impact on student retention.

The use of logistic regression was adopted to investigate whether national exam scores or secondary GPAs are better predictors of first-year retention in higher education [21]. High-stakes exams are entrance exams for higher education and are equivalent to national benchmark tests in South Africa. The study concluded by stating that school GPA predicts retention better in higher education compared to high-stakes national exams.

In another study, the authors assessed the performance of one deep learning algorithm and twenty supervised machine learning algorithms in predicting student retention [3]. All twenty-one algorithms were trained using the following variables: school accreditation, type of school, interest, average grades, gender, parent age, residence, parent salary, house area, parent's university attendance, and in-university retention. Random forest classifier, logistic regression CV, decision tree classifier, Nu support vector classifier (NuSVC), and linear support vector machine were amongst the twenty-one models used. Out of the twenty-one models used, the NuSVC algorithm emerged as the most effective machine learning method in predicting whether students would persist in their university enrollment or not.

Based on the literature reviewed, random forest and support vector machines were found to be the commonly used predictive models, mainly for classification tasks such as predicting whether a student will be retained (1) or not (0) in university. The most frequently used variables in predictive models for student retention mainly fell under two categories of student data: student demographics and academic performance. It is worth noting that all the models used predicted student retention on an individual student level, rather than an institutional level. As a result, the types of educational data analytics used are limited to learning analytics.

Chapter 3

Research Methodology

This chapter outlines the research method employed in this study. It includes a comprehensive step-by-step walk-through of the application of the Knowledge Discovery in Databases framework. Additionally, the chapter introduces the Bayesian Additive Regression Trees and Random Forest Regressor models, along with the accompanying metrics used to assess the predictive accuracy of these models in predicting student retention in tutorial classes. The process employed to transform attendance data into retention data is also explained.

3.1 Research Methods

This study employs a quantitative research method which incorporates the gathering of numerical secondary data and the use of mathematical, statistical, and computational methods to develop models. The quantitative research method draws its foundation from the positivism paradigm, which promotes the use of statistical analysis and various approaches that involve Bayesian inference, inferential statistics, probability theory, experimental design, as well as correlational and descriptive designs [1]. In this study, the Knowledge Discovery in Databases framework will be followed. The application of the KDD framework is widespread in the field of educational data mining and academic analytics research [22, 28]. The KDD framework provides a structured approach, comprising various steps, to convert raw data into actionable insights [14]. At an abstract level, KDD is concerned with developing methods for making sense of data [13]. The primary challenge addressed by the KDD framework is the transformation of low-level data into other forms that may be more compact, more abstract (such as a model of the data generation process), or more useful (for instance, a predictive model for estimating the value of future cases). This structured pathway facilitates the extraction of valuable patterns, trends, and knowledge from datasets, empowering informed decision-making. The KDD framework encompasses a series of essential steps: (1) data collection and understanding, (2) data pre-processing and transformation, (3) modeling, and (4) evaluation (as illustrated in Figure 3.1).

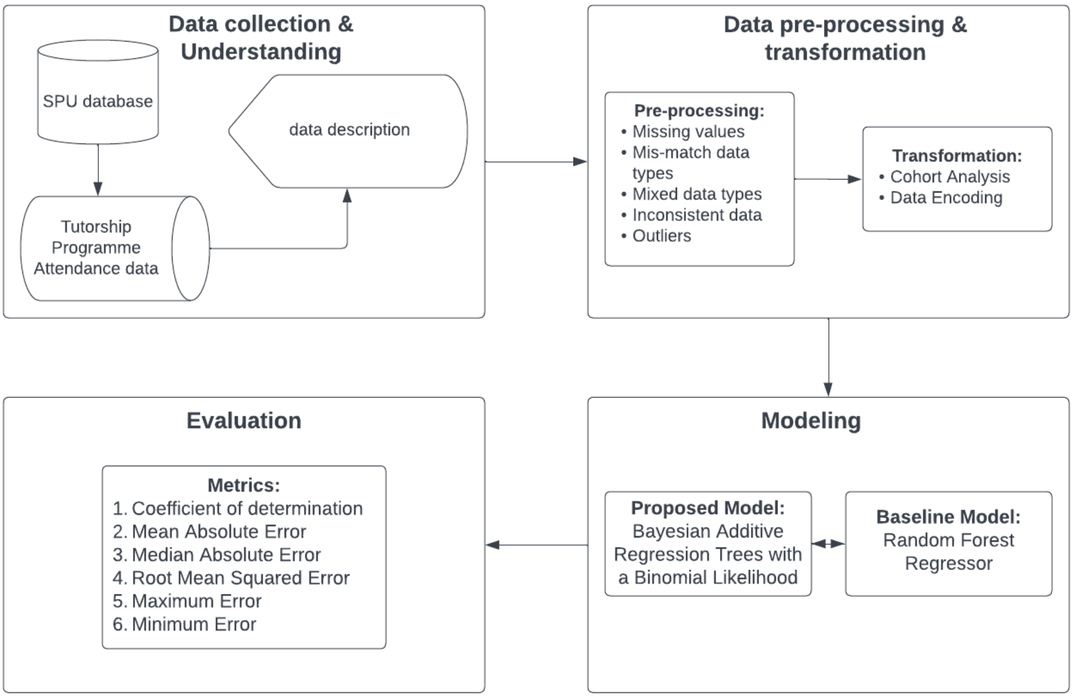


FIGURE 3.1: Knowledge discovery in database framework

3.1.1 Data Collection and Understanding

Prior to obtaining the secondary tutorship programme data, a data request letter was submitted to Sol Plaatje University. This letter provided information on the study's purpose, as well as the various measures implemented to mitigate potential risks and ensure the strict confidentiality of the acquired student data. It also provided insight into how the secondary data would be employed in the study (see data request letter on appendix A for further details).

The tutorship programme attendance data, sourced from the SPU database consisted of two central variables: tutorial date, and encoded (anonymized) student numbers. These two variables played a crucial role in deriving other variables used to predict attendance-based retention in tutorial classes. The tutorial date variable captured the dates on which students attended tutorial classes, thereby enabling the establishment of attendance patterns and trends. Conversely, the encoded student number variable provided encoded identifiers for each student, ensuring the anonymity and privacy of students represented in the attendance data. These encoded student numbers were instrumental in tracking student retention over the course of the tutorship programme (Table 3.1).

TABLE 3.1: Tutorship programme attendance data

Variable	**Description**
Tutorial date	Dates in which students attended tutorial classes
Encoded student number	Anonymized student number of students. Each anonymized student number serves as a unique identifier for each student

3.1.2 Data Pre-processing and Transformation

Data pre-processing and transformation is a key step that converts raw data into data that is more easily and effectively processed in models for more accurate and reliable results [28]. Firstly, data pre-processing was conducted to identify and handle missing data, mismatched data types, mixed data values, inconsistent data, and outliers. Secondly, data transformation was carried out through cohort analysis and data encoding. Cohort analysis is an analytical method that divides data into related groups called cohorts [20]. These cohorts share a common characteristic within a defined timespan. In this study cohorts were defined by the month in which students first started attending tutorial classes. The cohort analysis transformed TP attendance data to retention data. The transformed TP attendance data after cohort analysis consisted of 53 observations. Table 3.2 shows the description of variables derived from tutorial date, and encoded student numbers after cohort analysis.

Lastly, data encoding was applied to cohort and period as they were in date formats. The data encoding process involved transforming cohort and period into numeric formats that can be used as input in the modeling step.

TABLE 3.2: Tutorship programme attendance data after cohort analysis

Variable	**Description**
Cohort	The date students started attending tutorials.
Period	The date students stopped attending tutorials.
Cohort Age	The difference between the period and cohort, in days.
Students	The number of students that started attending tutorials for a particular cohort.
Active Students	The number of students currently attending tutorials at a particular period.
Retention	The number of active students divided by the number of students.

3.1.3 Modeling

In this study two models were implemented: the Random Forest Regressor (RFR) and Bayesian Addictive Regression Trees (BART). These models were implemented using historical TP attendance data to predict attendance-based retention in tutorial classes. The Random Forest Regressor was selected as the benchmark model to enable a robust performance comparison against the Bayesian Additive Regression Trees.

Random Forest Regressor

The RFR is a robust ensemble regression technique that leverages the combined power of multiple decision trees and employs a technique called bootstrapping and aggregation to improve predictive accuracy [6]. This technique provides several advantages, making it a valuable tool in modeling. From a computational perspective, the RFR offers several key strengths. It is known for its efficiency, as it is relatively fast both during the training phase and when making predictions. This speed is a result of its parallelizable nature, which allows for efficient implementation across high-dimensional datasets. The RFR depends on only one or two tuning parameters, which simplifies the modeling process. Additionally, it incorporates a built-in estimate of generalization error, aiding in the assessment of model performance and the prevention of overfitting [12].

The core principle of the RFR is to ensemble decision trees, combining their individual predictions to produce a more accurate and robust final output [31]. This is achieved through a process known as bootstrapping. When constructing the ensemble, the RFR algorithm repeatedly selects random samples with replacement from the original dataset. For each of these bootstrap samples, decision trees are trained to predict the response variables based on the corresponding variables. Specifically, for each iteration b (where b ranges from 1 to B, the number of bootstrapped samples), a random sample is drawn with replacement from the dataset (x, Y), yielding (x_b, Y_b). A decision tree regression model denoted as f_b is then trained on this sample. After completing the training phase for all B decision trees, the RFR is ready to make predictions. When presented with a new data point x' the ensemble regression model aggregates the predictions from all B individual decision trees to arrive at the final prediction. This aggregation is performed by calculating the average of the predictions made by each tree, represented as $\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$.

The RFR's strength lies in its ability to reduce overfitting, improve model robustness, and enhance predictive accuracy through the combination of multiple decision trees [28, 31]. The RFR model was constructed using the default RFR model parameters as specified by Sklearn (see appendix C Figure C.2).

Bayesian Additive Regression Trees

BART is a recent statistical approach that merges the principles of ensemble learning with nonparametric regression [11, 15, 17, 40]. What distinguishes BART is its construction within a Bayesian approach, enabling the quantification of prediction uncertainty through a model-based approach [15, 28, 37, 40]. BART's novelty lies in its capacity to adapt to complex relationships in the data while providing a robust and probabilistically grounded means of assessing predictive uncertainty, which sets it apart from traditional regression methods [40]. Mathematically, the BART model is represented as:

$$E[Y] = \phi\left(\sum_{j=0}^{m} g_j(X; T_j, M_j, \theta)\right) \tag{3.1}$$

Where X represents the model covariates (independent variables), each g_j is a tree of the form $g(X; T_j, M_j)$, where T_j represents the structure of a binary tree, i.e., the set of internal nodes and their associated decision rules, and a set of terminal nodes. While $M_j = \{\mu_{1,j}, \mu_{2,j}, \ldots, \mu_{b,j}\}$ represents the values at the b_j terminal nodes, ϕ represents an arbitrary probability distribution that will be used as the likelihood in the model, and θ represents other parameters not modeled as a sum of trees [19].

In this study the BART model was specified as follows:

$$N_{\text{active students}} \sim \text{Bin}(N_{\text{students}}, p) \tag{3.2}$$

$$\text{logit}(p) = \text{BART}(\text{cohort age}, \text{month}) \tag{3.3}$$

Where $\text{Bin}(N_{\text{students}}, p)$ represents the likelihood probability distribution for the number of active students. This likelihood was used to indicate that the number of active students follows a binomial distribution. The selection of a Binomial likelihood in the BART model was motivated by its suitability for count data, reflecting the act of counting active students within a group of students. Here, p represents the retention rate. In the BART model, the logit function given by $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ was used as a transformation function to map the retention rate to the range $(-\infty, +\infty)$ so that the range is not constrained to $(0, 1)$. The logit transformation allowed for a flexible and non-linear estimation of log-odds of success, accounting for the complex interactions between cohort age and month. To interpret the results in terms of retention rate $(0, 1)$, the inverse logit function, $p = \frac{\exp(\text{logit})}{1+\exp(\text{logit})}$, was used, where logit represents the log-odds value. This transformation allowed the conversion of the model's log-odds back into the $(0, 1)$ range, enabling the estimation of a retention rate at a given time point (see appendix C

Figure C.1 for BART model code implementation).

Figure 3.2 shows the graphical representation of the BART model with a binomial likelihood. The BART model was fitted using the default pymc Markov chain Monte Carlo (MCMC) algorithm to generate 2000 samples of all model parameters and predictions from the corresponding posterior probability distributions.

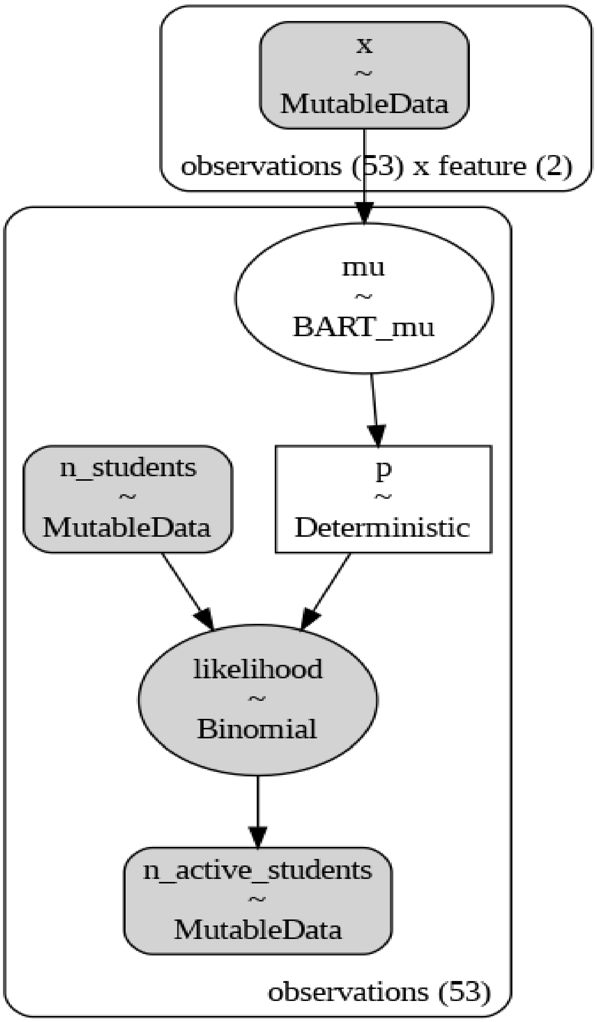


FIGURE 3.2: Graphical representation of the BART model with a binomial likelihood

3.1.4 Evaluation

The retention variable in tutorial classes was modeled as a continuous variable that ranges between 0 and 1, where 0 means 0% of the students are retained in tutorial classes, while 1 means 100% of the students are retained in tutorial classes. The retention prediction error was evaluated using six metrices that are commonly used for continuous variables, namely minimum error, maximum error, mean absolute error (MAE), median absolute error (MedAE), root mean squared error (RMSE), and coefficient of determination. Each evaluation metric captures different aspects of model performance. By using all six of these metrics, a more comprehensive assessment of model performance was provided. This allowed the analysis of various facets such as the range of errors (minimum and maximum), average errors (MAE), robustness to outliers (MedAE), precision (RMSE), and the proportion of variance explained (coefficient of determination). In the error metric calculations, $\hat{y}_i$ represents the predicted value of the $i - th$ sample and y_i is the corresponding true value.

Coefficient of Determination (R-squared Score)

The R-squared was used to quantify the proportion of variance in the retention variable explained by the model. It ranges from 0 to 1, where a higher value indicates a better fit. R-squared of 1 means the model perfectly predicts the retention, while 0 means the model fails to explain any variation [10, 30].

The estimated R^2 is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.4}$$

where $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Mean Absolute Error

The MAE was used to calculate the average absolute difference between the predicted retention and actual retention values. The MAE provides a measure of the average magnitude of errors, regardless of their direction. Lower MAE indicates better accuracy and closer predictions to the ground truth [16, 30].

The MAE estimated over n_{samples} is defined as:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}-1}} |y_i - \hat{y}_i| \tag{3.5}$$

Median Absolute Error

Like MAE, the median absolute error was used to provide a measure of the average magnitude of errors. However, instead of averaging the errors, the median value was considered. The MedAE is less sensitive to outliers compared to MAE, making it a robust metric [30].

The MedAE estimated over n_{samples} is defined as:

$$\text{MedAE}(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \tag{3.6}$$

Root Mean Squared Error

The RMSE was used to calculate the square root measure of the squared differences between the predicted retention and actual retention values. RMSE penalizes larger error values. A lower RMSE indicates better precision [16, 30].

The RMSE estimated over n_{samples} is defined as:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}-1}} (y_i - \hat{y}_i)^2} \tag{3.7}$$

Maximum Error

The maximum error was calculated to represent the largest deviation between the predicted retention values and actual retention values. It was used to capture the worst-case error between the predicted retention value and the true retention value [30].

The maximum error is defined as:

$$\text{Max Error}(y, \hat{y}) = \max(|y_i - \hat{y}_i|) \tag{3.8}$$

Minimum Error

The minimum error was calculated to represent the smallest deviation between the predicted retention values and actual retention values.

The minimum error is defined as:

$$\text{Min Error}(y, \hat{y}) = \min(|y_i - \hat{y}_i|) \tag{3.9}$$

Chapter 4

Results and Discussion

The findings in this chapter are structured into sections covering descriptive analysis, model evaluation, BART model highest density Interval estimates, and tutorial attendance dashboard. Initially, the chapter unveils descriptive analysis results, featuring a statistical summary of tutorial class attendance data. This is followed by a time series analysis of student retention rates. Subsequently, the chapter delves into the presentation of model performance results. Lastly, it demonstrates the practical application of the highest density interval for uncertainty quantification and introduces the tutorial class attendance dashboard.

4.1 Descriptive Analysis

After applying the methods for data preprocessing and transformation described in section 3.1.2, six variables were derived: ’Cohort’, ’Period’, ’Cohort Age’, ’Students’, ’Active Students’, and ’Retention’.

Table 4.1 presents a comprehensive overview of key descriptive statistics for the six variables of interest. The ’Cohort’ and ’Period’ variables, representing the start and end dates of student attendance, do not have meaningful measures like means, medians, or standard deviations due to their date nature; however, they provide a range, with ’Cohort’ spanning from 1 January 2022 to 1 October 2022, and ’Period’ ranging from 1 March 2022 to 1 December 2022. ’Cohort Age’ has a mean of 141.62 days, indicating that, on average, students attended tutorials for this duration. The median of 122 days shows the typical duration, while a standard deviation of 85.69 days reflects some variation. On average 183.60 ’Students’ are present in tutorial classes in any given day, with considerable variability (standard deviation of 155.08) between cohorts, ranging from a minimum of 10 to a maximum of 421 students. ′Active Students′ has an average of 50.87 and a median of 32, indicating the typical number of students actively participating, with significant variability (standard deviation of 65.17). 'Retention' showcases an average retention rate of 33.34%, a typical rate of 24.23%, and a standard deviation of 28.54%, reflecting the diversity in how well students are retained, with rates ranging from a minimum of 0.45% to a maximum of 94.92%. These statistics provide valuable insights into the dynamics of student participation and retention in this study.

TABLE 4.1: Descriptive statistics of variables of interest

Variable	**Mean**	**Median**	**Std. Dev.**	**Max**	**Min**
Cohort	-	-	-	2022-10-01	2022-01-01
Period	-	-	-	2022-12-01	2022-03-01
Cohort Age	141.6226	122	85.6869	344	28
Students	183.6038	80	155.0793	421	10
Active Students	50.8679	32	65.1652	263	1
Retention	33.34 %	24.23 %	28.54 %	94.92 %	0.45 %

Figure 4.1 displays the variation in retention rates over time across nine different cohorts. Each cohort indicates the starting date of students attending tutorial classes. The retention rate is at its highest when students commence their tutorial classes and gradually decreases until June (2022-06) and July (2022-07), after which it increases again and then decreases towards November (2022-11) and December (2022-12). Notably, the June, July, November, and December period coincides with mid-year and end-year exams and the semester break, during which the retention rate is at its lowest, suggesting that students discontinue attending tutorial classes to focus on exam preparation with a holiday break happening after. This illustrates a clear seasonality component in the retention, as depicted in Figure 4.1, where seasonality peaks decrease over time for each cohort of students.

The presence of seasonality in student retention rates highlights its nonlinearity.

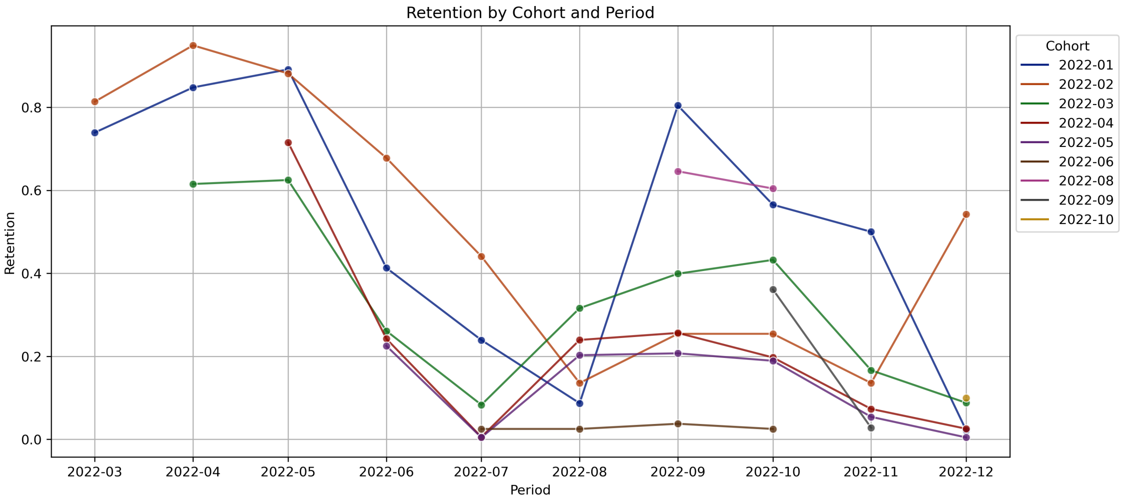


FIGURE 4.1: Retention by cohort and period

This nonlinearity is evident in the fluctuating retention rates across various periods, driven by factors like exam periods and holiday seasons. Using nonlinear models, such as BART and RFR, is crucial for capturing these non-linear patterns in student retention rates. These models offer the necessary flexibility to capture complex relationships, handle interactions, and effectively adapt to the nonlinear patterns within the data [40].

4.2 Model Evaluation

The BART model with a Binomial likelihood is assessed in comparison to the RFR using six key metrics: R-squared, MAE, RMSE, MedAE, Max Error, and Min Error. Student retention, ranging from 0% to 100%, serves as a critical indicator, representing the percentage of students retained in tutorial classes.

Table 4.2 displays the results for BART with a Binomial likelihood and the RFR model. BART demonstrates strong predictive capabilities, outperforming RFR across all key evaluation metrics. With a significantly higher R^2 score of 0.9414 compared to RFR's 0.9150, indicating that BART effectively captures a greater proportion of the variance in student retention. Additionally, BART yields a lower MAE of 4.75% as opposed to RFR's 6.66%, indicating more accurate predictions on average. The RMSE for BART (6.85%) is also lower than that of RFR (8.25%), signifying that its predictions are generally more precise. The MedAE of 3% for BART reflects its consistency in providing predictions close to the actual retention, while RFR shows a MedAE of 6%. Furthermore, BART achieves a slightly lower maximum error (19%) compared to RFR (20%). Both models exhibit a minimum error of 0%, indicating accurate predictions in some instances.

TABLE 4.2: BART and RFR model evaluation

Model	R^2 Score	MAE	RMSE	MedAE	Max Error	Min Error
BART	0.9414	4.75 %	6.85 %	3 %	19 %	0 %
RFR	0.9150	6.66 %	8.25 %	6 %	20 %	0 %

The BART model demonstrates stronger predictive capabilities in predicting student retention in tutorial classes as compared to RFR, this would make BART a preferred choice for educational institutions in need of robust predictive capabilities.

4.3 BART Model Highest Density Interval Estimates

Figure 4.2 and Figure 4.3 show the 94% highest density interval (HDI) uncertainty estimates for a set of individual cohorts. The HDI is a range of values that captures a certain percentage of a model's parameters [19]. It provides a measure of the uncertainty in the parameter's value and can be used to make inferences about the parameter [19]. The 94% HDI in Figure 4.2 and Figure 4.3 is a range that captures 94% of the posterior distribution of BART parameters. This means that there is a 94% probability that the true retention rate falls within this interval. A wide HDI interval is an indication of great uncertainty, while a narrow HDI interval is an indication of great certainty. Narrower HDIs indicate more reliable predictions as they suggest the model has effectively minimized uncertainty. This precision leads to a higher level of confidence in the prediction, offering a more trustworthy basis for decision-making. Precision in the context of HDIs refers to the narrowness of the interval that captures the range of plausible values for a prediction [27].

In Figure 4.2 and Figure 4.3, the 94% HDI interval is wide for cohorts 2022-01, 2022-02, 2022-06, 2022-08, and 2022-09, indicating great uncertainty in the BART model's predictions for student retention in these cohorts. Conversely, the 94% interval is narrow for cohorts 2022-03, 2022-04, and 2022-05, indicating a great level of certainty and reliability in the BART model's predictions for student retention in these cohorts.

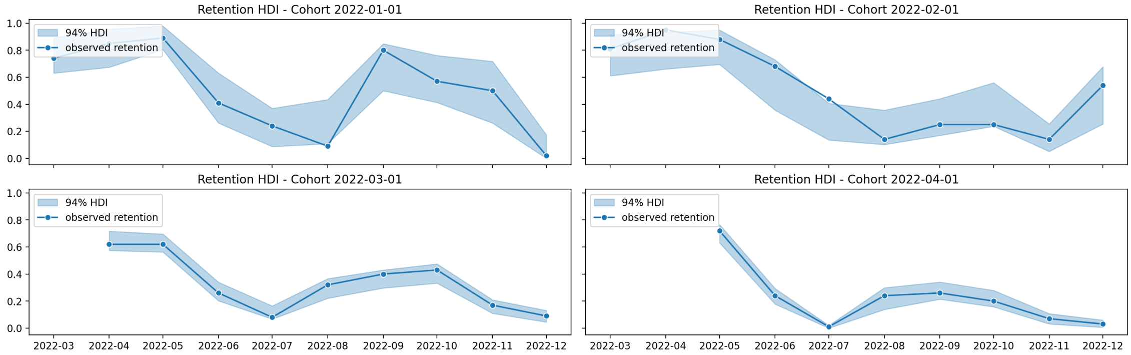


FIGURE 4.2: BART 94% HDI for 2022-01 to 2022-04 cohorts

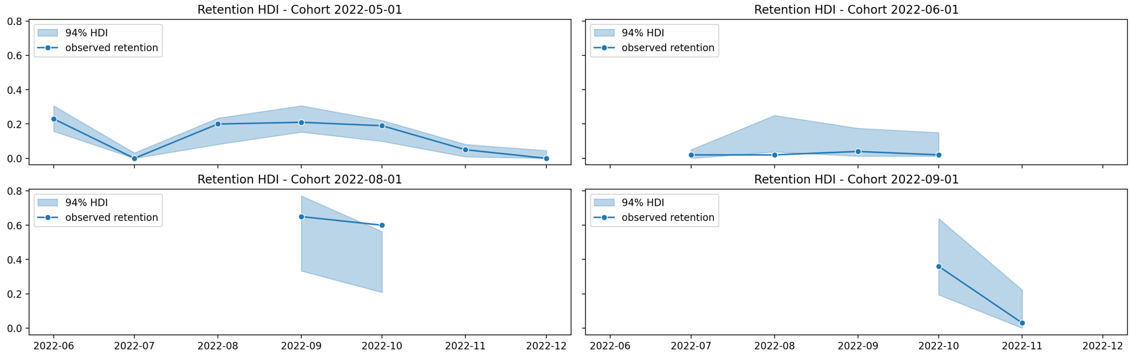


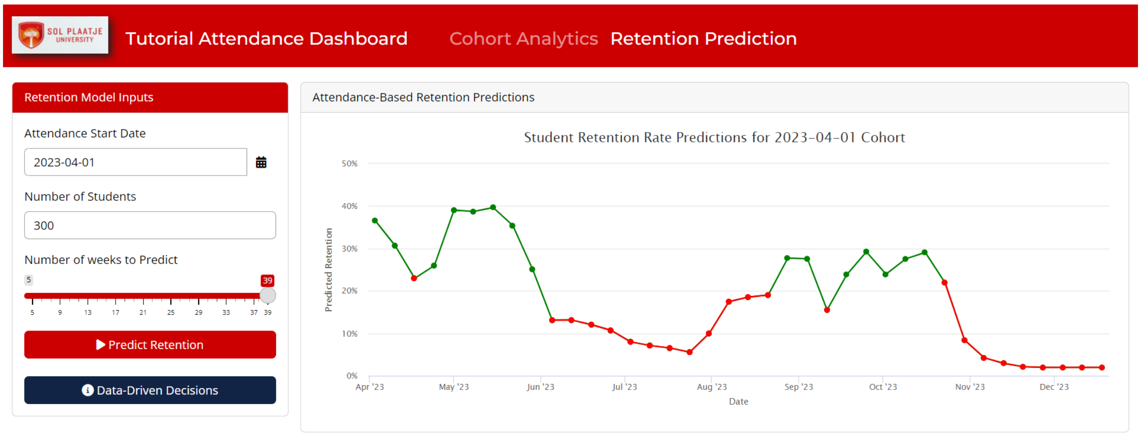
FIGURE 4.3: BART 94% HDI for 2022-05 to 2022-09 cohorts

The high degree of certainty observed in cohorts 2022-03 to 2022-05, as evident by the 94% HDI closely aligning with the observed retention rate, underscores the BART model's accuracy in predicting student retentions from the onset of tutorials to just before the start of exams. In situations requiring a single prediction (point estimate), such a prediction can be obtained by calculating the average or median of the predicted retention values within the 94% HDI.

4.4 Tutorial Class Attendance Dashboard

One significant contribution was the development of a dashboard tailored for the CTLPD. This dashboard serves the dual purpose of predicting future attendance-based retention rates and offering detailed insights into TP attendance-based retention rates. The dashboard was developed using 'Shiny', a web application framework with interactive analytical capabilities [29].

Within the dashboard, the 'Retention Prediction' tab shown in Figure 4.4 hosts an intuitive 'Retention Model Inputs' section. This section enables education staff to input crucial data points, including the start date of student attendance in tutorial classes, the current count of attending students, and the desired projection period in weeks for attendance-based retention rates. Additionally, the inclusion of a 'Data-driven Decisions' button provides guidance on using the attendance-based retention predictive model for informed insights (refer to appendix D for detailed instructions). The accompanying line chart, positioned on the right, visualizes predicted attendance-based retention rates for a user-defined duration. The chart distinguishes future retention rates surpassing the institution's average (depicted in green) from those falling below it (depicted in red).


Retention Model Inputs

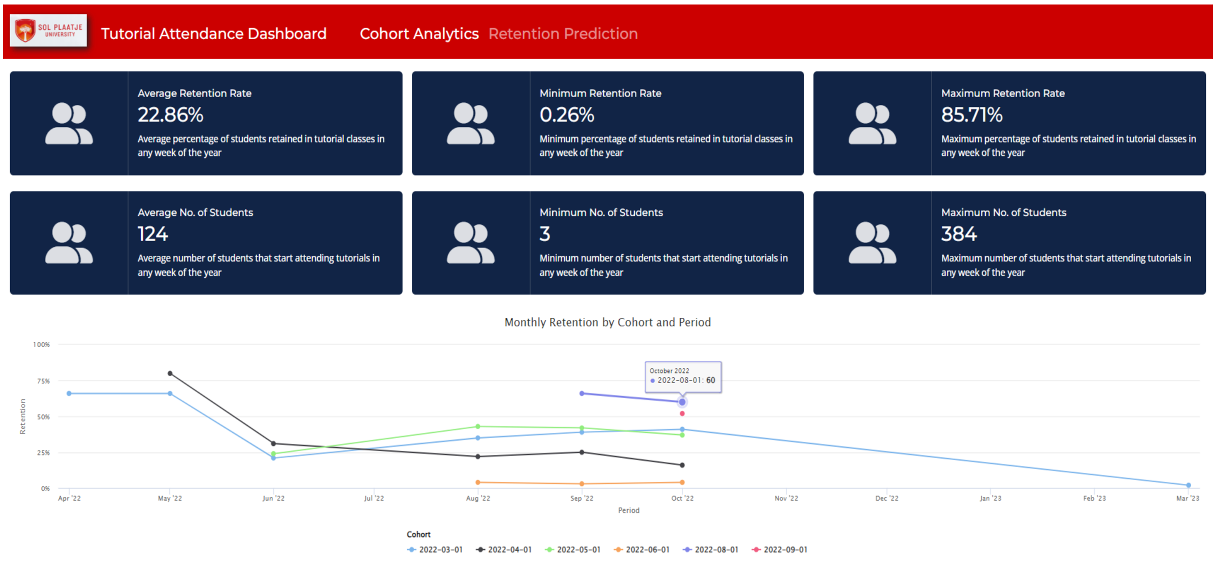
Attendance Start Date: 2023-04-01

Number of Students: 300

Number of weeks to Predict: 5 – 39

FIGURE 4.4: Retention prediction dashboard tab

Moreover, the 'Cohort Analytics' tab in Figure 4.5 provides in-depth descriptive analytics pertaining to different student cohorts starting tutorial classes in specific months and weeks throughout the year. These analytical insights have the same interpretation as the descriptive analysis provided in section 4.1, serving to outline and summarize attendance and retention patterns within tutorial classes.



Average Retention Rate
22.86%
Average percentage of students retained in tutorial classes in any week of the year

Minimum Retention Rate
0.26%
Minimum percentage of students retained in tutorial classes in any week of the year

Maximum Retention Rate
85.71%
Maximum percentage of students retained in tutorial classes in any week of the year

Average No. of Students
124
Average number of students that start attending tutorials in any week of the year

Minimum No. of Students
3
Minimum number of students that start attending tutorials in any week of the year

Maximum No. of Students
384
Maximum number of students that start attending tutorials in any week of the year

Monthly Retention by Cohort and Period

FIGURE 4.5: Cohort analytics dashboard tab

Chapter 5

Conclusions

This chapter is split into two sections. The first section addresses the research questions, offering conclusive insights. The second section directs attention to potential for future research, outlining ways in which this study can be extended.

5.1 Answers to Research Questions

RQ 1: To what extent does the performance of a Bayesian model differ from that of a non-Bayesian model in terms of predicting retention in tutorial classes, as evaluated using metrics including minimum error, maximum error, mean absolute error, median absolute error, root mean squared error, and coefficient of determination?

On average, the Bayesian Additive Regression Tree model outperformed the Random Forest Regressor model by approximately 20% when predicting retention in tutorial classes across five error metrics, including minimum error, maximum error, mean absolute error, median absolute error, and root mean squared error. In specific terms, the BART model exhibits an average prediction error of 6.72% when predicting retention rates on a 0 to 100% scale, while the RFR model displays an average error of 8.182%. Furthermore, BART achieves a higher coefficient of determination (0.9414) compared to RFR's (0.9150), indicating that BART effectively captures a greater proportion of the variance in student retention.

While the difference in error metrics between BART and RFR may not be substantial, BART demonstrates stronger predictive capabilities in estimating student retention. These performance distinctions stem from the fundamental differences in their methodologies. BART operates within a Bayesian framework, accounting for parameter uncertainties and employing probabilistic models for each decision tree. In contrast, the Random Forest Regressor relies on creating an ensemble of decision trees through bootstrapping and aggregation without explicit probabilistic modeling. BART also offers more flexibility in modeling complex relationships but can be computationally more demanding than Random Forest Regressor. Additionally, BART may involve more tuning parameters than the user-friendly and straightforward Random Forest Regressor. These distinctions emphasize the variations in their underlying methodologies and suitability for specific applications. When choosing a model for predicting retention in tutorial classes, it is essential to consider these differences, even though their performance differences may appear minor.

RQ 2: How can the highest density interval be used to summarize the uncertainty in retention predictions and make inferences about retention in tutorial classes?

The Highest Density Interval can be used to obtain the worst-case and best-case scenarios, as the lower and upper bounds of the HDI can be considered as the worst-case and best-case when predicting student retention rates. Large differences between the lower bound and upper bounds of the HDI can be used to signify great uncertainty while small differences can be used to signify great certainty.

The lower bound value of the HDI can be used to identify the minimum level of support needed to maintain student retention at a reasonable level, while the upper bound value can be used to determine the maximum impact that a support programme is likely to have on student retention in tutorial classes.

5.2 Future Work

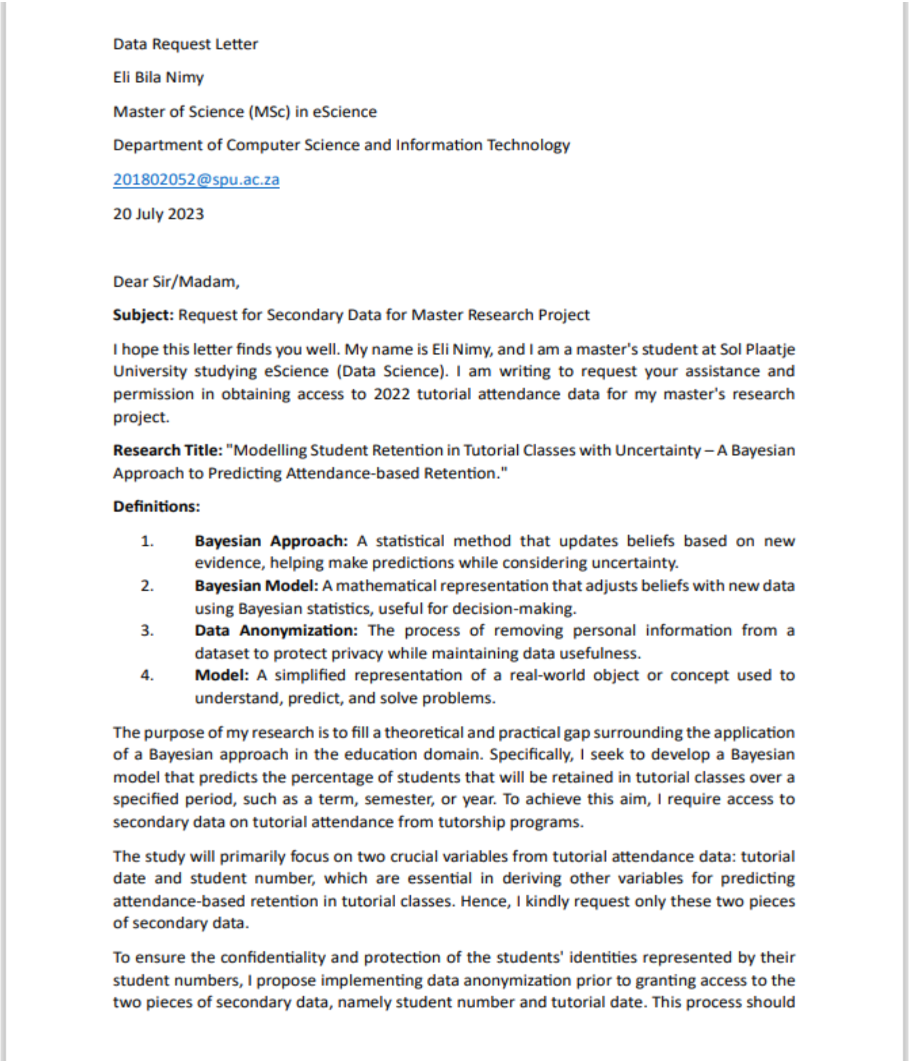
This study mainly focused on the development of a single Bayesian model to predict retention in tutorial classes using derived variables from cohort analysis. Tutorship support is one of the various student support programmes that institutions have. Future work can extend this study by exploring the development and implementation of different machine learning and statistical models for predicting student retention within different student support programmes for timely interventions. Furthermore, future studies can focus on incorporating a broader range of student data variables with the application of interpretable machine learning techniques to explain key factors that influence student retention within student support programmes.

The Bayesian approach is one of the established methods for achieving uncertainty quantification in the context of model parameters and predictions. In addition to Bayesian methods, Conformal Prediction (CP) represents another robust statistical framework for quantifying uncertainty. CP excels in generating prediction regions that capture the inherent variability of point predictions, adding an essential layer of reliability to predictive modeling [2]. Leveraging CP within the context of student retention predictive models can provide a means to enhance the reliability and precision of machine learning model predictions.

Finally, the importance and relevance of this study extends to various stakeholders in the educational ecosystem. Educational institutions, administrators, and policymakers can benefit from the research findings by gaining insights into how future tutorship programme student retention rates can be predicted. This information can help various stakeholders in the educational ecosystem develop tailored intervention strategies to improve student retention. Tutorship programme coordinators can use the predictive model(s) to identify periods of low or high retention and implement timely interventions to encourage engagement accordingly. Moreover, the foresight provided by the expected student retentions can assist in strategic resource allocation, enabling more informed planning and budgeting for tutorship support programmes.

Appendix A

Data Request Letter



Data Request Letter

Eli Bila Nimy

Master of Science (MSc) in eScience

Department of Computer Science and Information Technology

201802052@spu.ac.za

20 July 2023

Dear Sir/Madam,

Subject: Request for Secondary Data for Master Research Project

I hope this letter finds you well. My name is Eli Nimy, and I am a master's student at Sol Plaatje University studying eScience (Data Science). I am writing to request your assistance and permission in obtaining access to 2022 tutorial attendance data for my master's research project.

Research Title: "Modelling Student Retention in Tutorial Classes with Uncertainty – A Bayesian Approach to Predicting Attendance-based Retention."

Definitions:

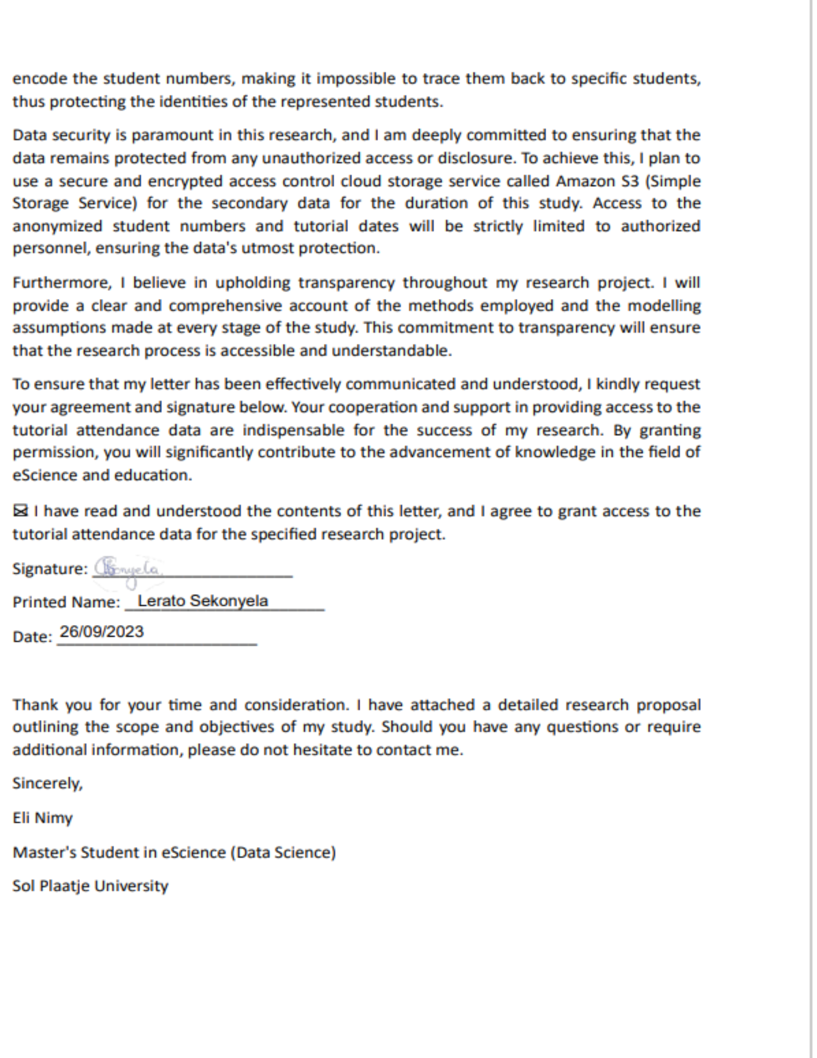
1. **Bayesian Approach:** A statistical method that updates beliefs based on new evidence, helping make predictions while considering uncertainty.
2. **Bayesian Model:** A mathematical representation that adjusts beliefs with new data using Bayesian statistics, useful for decision-making.
3. **Data Anonymization:** The process of removing personal information from a dataset to protect privacy while maintaining data usefulness.
4. **Model:** A simplified representation of a real-world object or concept used to understand, predict, and solve problems.

The purpose of my research is to fill a theoretical and practical gap surrounding the application of a Bayesian approach in the education domain. Specifically, I seek to develop a Bayesian model that predicts the percentage of students that will be retained in tutorial classes over a specified period, such as a term, semester, or year. To achieve this aim, I require access to secondary data on tutorial attendance from tutorship programs.

The study will primarily focus on two crucial variables from tutorial attendance data: tutorial date and student number, which are essential in deriving other variables for predicting attendance-based retention in tutorial classes. Hence, I kindly request only these two pieces of secondary data.

To ensure the confidentiality and protection of the students' identities represented by their student numbers, I propose implementing data anonymization prior to granting access to the two pieces of secondary data, namely student number and tutorial date. This process should

FIGURE A.1: Data request letter page 1



encode the student numbers, making it impossible to trace them back to specific students, thus protecting the identities of the represented students.

Data security is paramount in this research, and I am deeply committed to ensuring that the data remains protected from any unauthorized access or disclosure. To achieve this, I plan to use a secure and encrypted access control cloud storage service called Amazon S3 (Simple Storage Service) for the secondary data for the duration of this study. Access to the anonymized student numbers and tutorial dates will be strictly limited to authorized personnel, ensuring the data's utmost protection.

Furthermore, I believe in upholding transparency throughout my research project. I will provide a clear and comprehensive account of the methods employed and the modelling assumptions made at every stage of the study. This commitment to transparency will ensure that the research process is accessible and understandable.

To ensure that my letter has been effectively communicated and understood, I kindly request your agreement and signature below. Your cooperation and support in providing access to the tutorial attendance data are indispensable for the success of my research. By granting permission, you will significantly contribute to the advancement of knowledge in the field of eScience and education.

☒ I have read and understood the contents of this letter, and I agree to grant access to the tutorial attendance data for the specified research project.

Signature: ____________________

Printed Name: Lerato Sekonyela

Date: 26/09/2023

Thank you for your time and consideration. I have attached a detailed research proposal outlining the scope and objectives of my study. Should you have any questions or require additional information, please do not hesitate to contact me.

Sincerely,

Eli Nimy

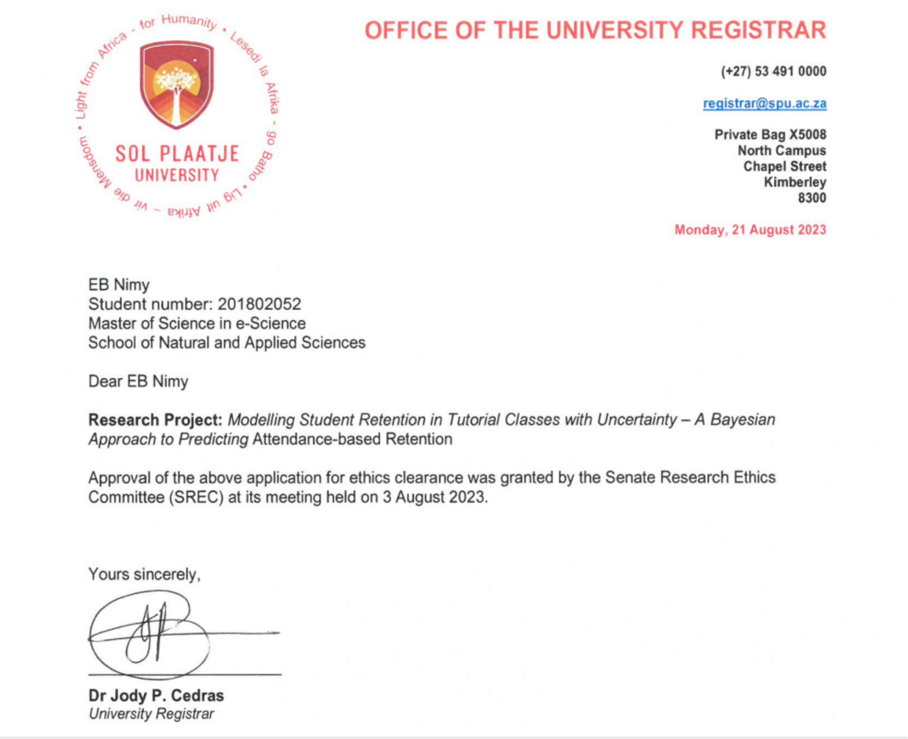
Master's Student in eScience (Data Science)

Sol Plaatje University

FIGURE A.2: Data request letter page 2

Appendix B

Ethical Clearance



SOL PLAATJE UNIVERSITY

Light from Africa - for Humanity • Lesedi la Afrika - go Batho • Lig uit Afrika ~ vir die Mensdom •

OFFICE OF THE UNIVERSITY REGISTRAR

(+27) 53 491 0000

registrar@spu.ac.za

Private Bag X5008
North Campus
Chapel Street
Kimberley
8300

Monday, 21 August 2023

EB Nimy
Student number: 201802052
Master of Science in e-Science
School of Natural and Applied Sciences

Dear EB Nimy

Research Project: *Modelling Student Retention in Tutorial Classes with Uncertainty – A Bayesian Approach to Predicting* Attendance-based Retention

Approval of the above application for ethics clearance was granted by the Senate Research Ethics Committee (SREC) at its meeting held on 3 August 2023.

Yours sincerely,

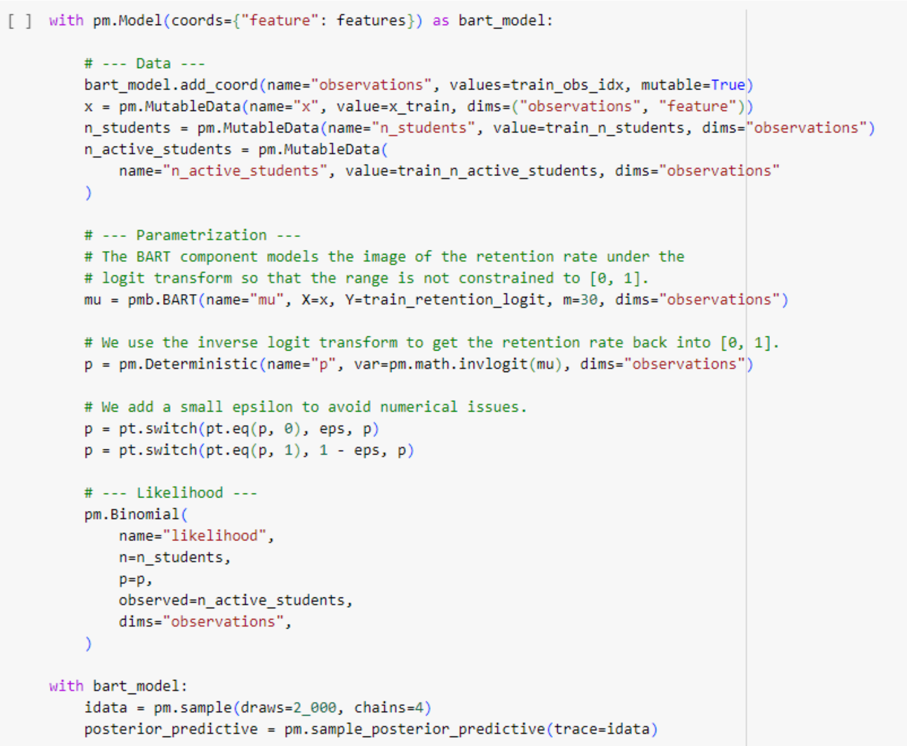
Dr Jody P. Cedras
University Registrar

FIGURE B.1: Ethical clearance letter

Appendix C

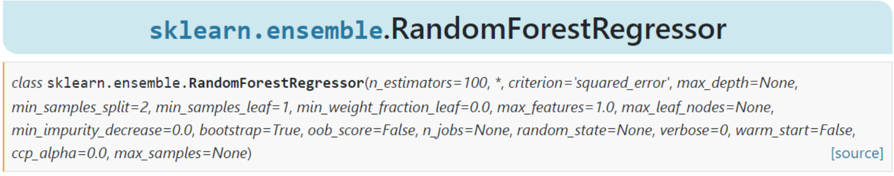
Model Specifications and Predictive Results

C.1 Model Specification



```
[ ]  with pm.Model(coords={"feature": features}) as bart_model:

         # --- Data ---
         bart_model.add_coord(name="observations", values=train_obs_idx, mutable=True)
         x = pm.MutableData(name="x", value=x_train, dims=("observations", "feature"))
         n_students = pm.MutableData(name="n_students", value=train_n_students, dims="observations")
         n_active_students = pm.MutableData(
             name="n_active_students", value=train_n_active_students, dims="observations"
         )

         # --- Parametrization ---
         # The BART component models the image of the retention rate under the
         # logit transform so that the range is not constrained to [0, 1].
         mu = pmb.BART(name="mu", X=x, Y=train_retention_logit, m=30, dims="observations")

         # We use the inverse logit transform to get the retention rate back into [0, 1].
         p = pm.Deterministic(name="p", var=pm.math.invlogit(mu), dims="observations")

         # We add a small epsilon to avoid numerical issues.
         p = pt.switch(pt.eq(p, 0), eps, p)
         p = pt.switch(pt.eq(p, 1), 1 - eps, p)

         # --- Likelihood ---
         pm.Binomial(
             name="likelihood",
             n=n_students,
             p=p,
             observed=n_active_students,
             dims="observations",
         )

     with bart_model:
         idata = pm.sample(draws=2_000, chains=4)
         posterior_predictive = pm.sample_posterior_predictive(trace=idata)
```

FIGURE C.1: Binomial likelihood Bayesian additive regression trees model code



sklearn.ensemble.RandomForestRegressor

class `sklearn.ensemble.RandomForestRegressor`(*n_estimators=100, *, criterion='squared_error', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=1.0, max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, ccp_alpha=0.0, max_samples=None*) [source]

FIGURE C.2: Random forest regressor default Sklearn model

C.2 BART Model Posterior Predictive Results

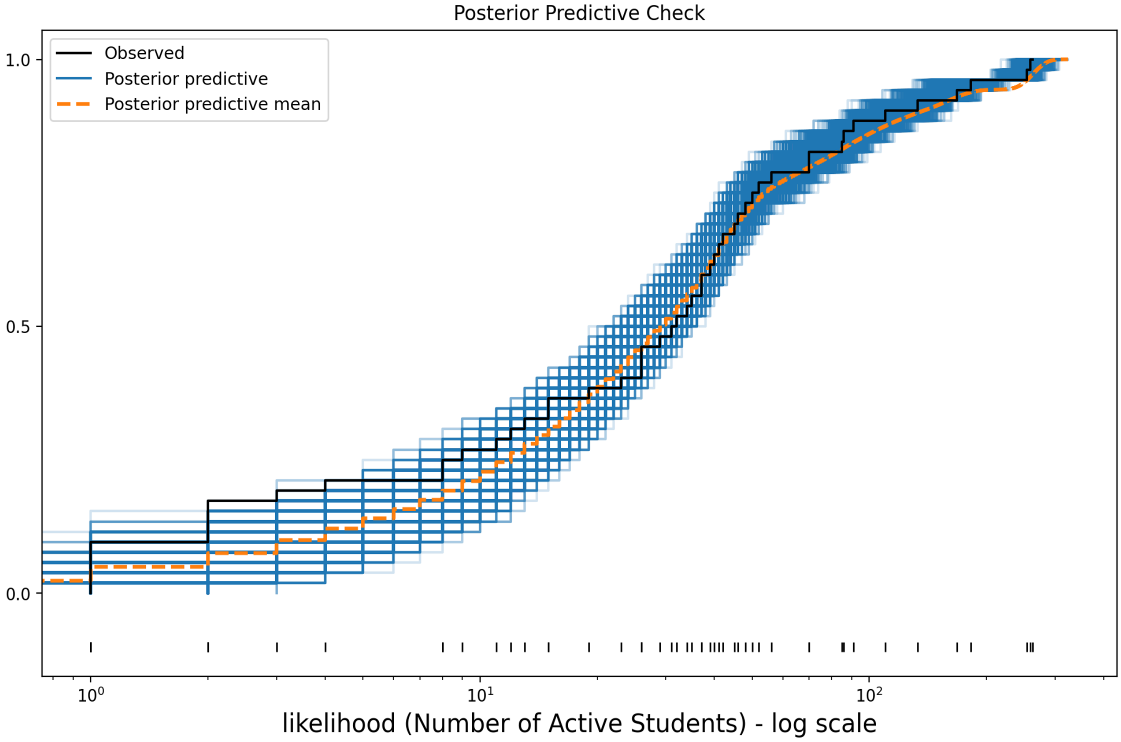


FIGURE C.3: BART model posterior predictive check

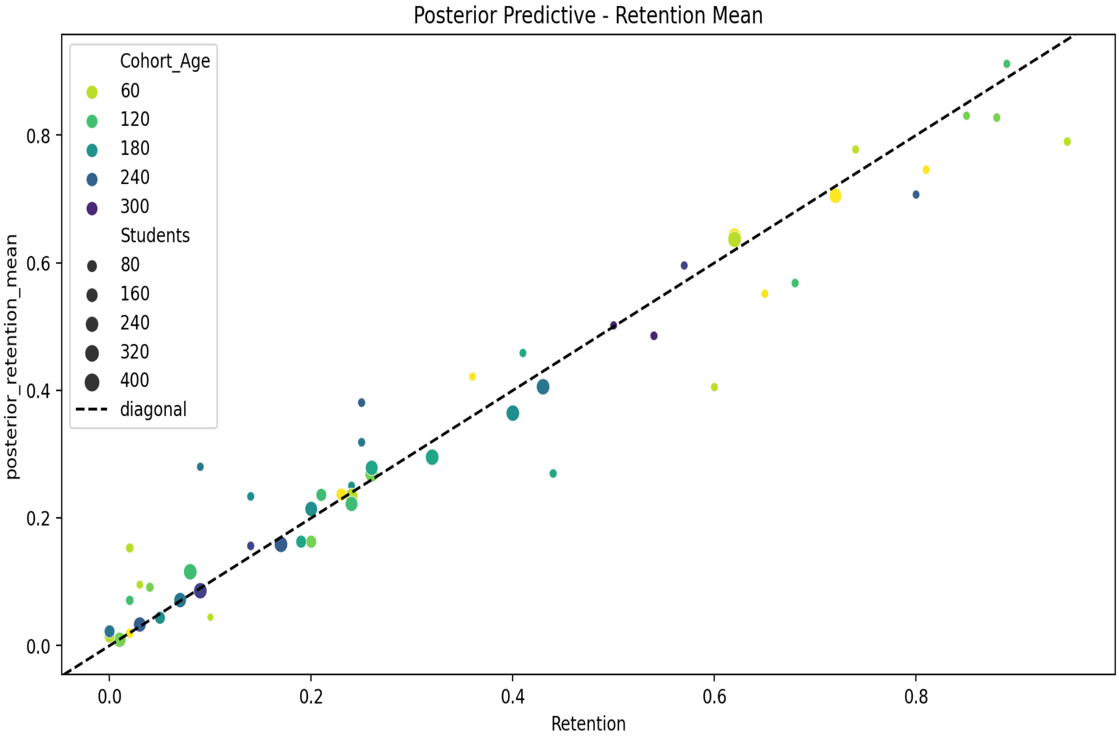
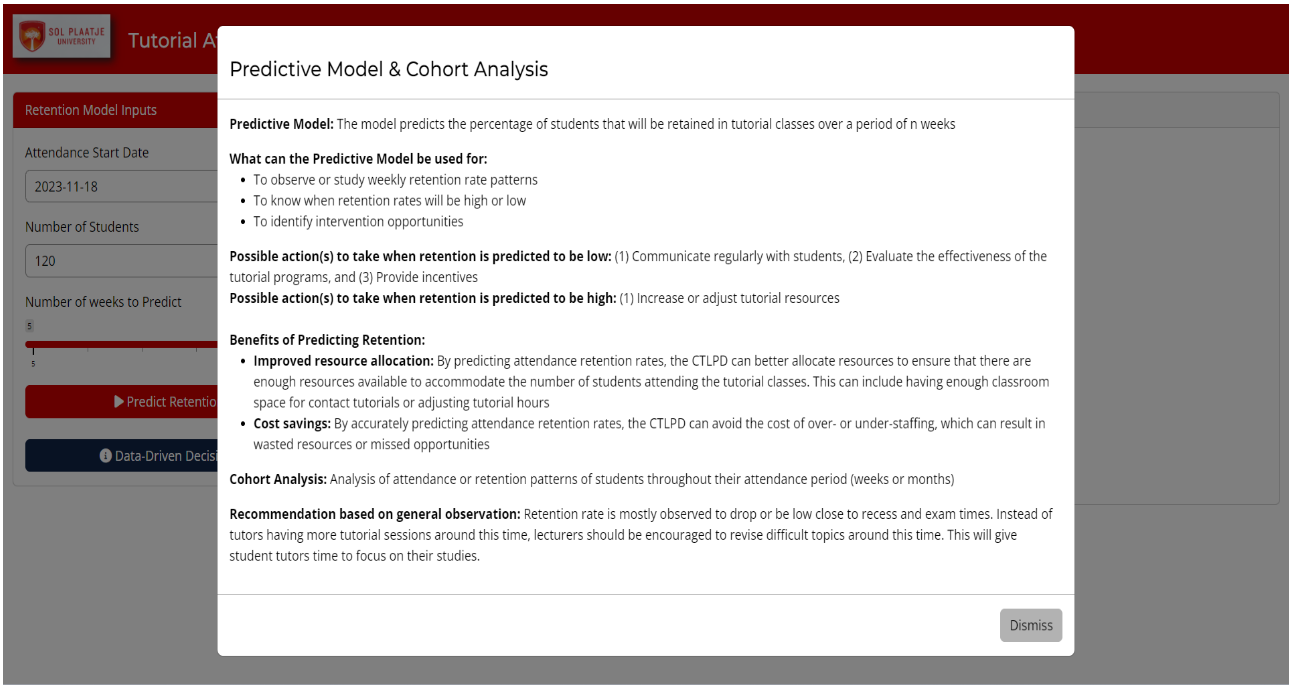


FIGURE C.4: BART model posterior predictive mean

Appendix D

Tutorial Attendance Dashboard Modal



Retention Model Inputs

Attendance Start Date

2023-11-18

Number of Students

120

Number of weeks to Predict

5

Predictive Model & Cohort Analysis

Predictive Model: The model predicts the percentage of students that will be retained in tutorial classes over a period of n weeks

What can the Predictive Model be used for:

- To observe or study weekly retention rate patterns
- To know when retention rates will be high or low
- To identify intervention opportunities

Possible action(s) to take when retention is predicted to be low: (1) Communicate regularly with students, (2) Evaluate the effectiveness of the tutorial programs, and (3) Provide incentives
Possible action(s) to take when retention is predicted to be high: (1) Increase or adjust tutorial resources

Benefits of Predicting Retention:

- **Improved resource allocation:** By predicting attendance retention rates, the CTLPD can better allocate resources to ensure that there are enough resources available to accommodate the number of students attending the tutorial classes. This can include having enough classroom space for contact tutorials or adjusting tutorial hours
- **Cost savings:** By accurately predicting attendance retention rates, the CTLPD can avoid the cost of over- or under-staffing, which can result in wasted resources or missed opportunities

Cohort Analysis: Analysis of attendance or retention patterns of students throughout their attendance period (weeks or months)

Recommendation based on general observation: Retention rate is mostly observed to drop or be low close to recess and exam times. Instead of tutors having more tutorial sessions around this time, lecturers should be encouraged to revise difficult topics around this time. This will give student tutors time to focus on their studies.

FIGURE D.1: Tutorial attendance dashboard data-driven decision modal

Bibliography

[1] O.B. Adedoyin. *Qualitative Research Method*. Accessed: 10 April 2023. 2020. URL: https://www.researchgate.net/publication/340594619_Quantitative_Research_Method.

[2] Anastasios N Angelopoulos and Stephen Bates. “A gentle introduction to conformal prediction and distribution-free uncertainty quantification”. In: *arXiv preprint arXiv:2107.07511* (2021).

[3] Samer M Arqawi et al. “Predicting university student retention using artificial intelligence”. In: *International Journal of Advanced Computer Science and Applications* 13.9 (2022).

[4] RSJD Baker et al. “Data mining for education”. In: *International encyclopedia of education* 7.3 (2010), pp. 112–118.

[5] Roberto Bertolini, Stephen J Finch, and Ross H Nehm. “An application of Bayesian inference to examine student retention and attrition in the STEM classroom”. In: *Frontiers in Education*. Vol. 8. Frontiers Media SA. 2023, p. 1073829.

[6] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[7] Barbara Flores Caballero. “Higher Education: Factors and Strategies for Student Retention.” In: *HETS Online Journal* 10 (2020).

[8] John P Campbell, Peter B DeBlois, and Diana G Oblinger. “Academic analytics: A new tool for a new era”. In: *EDUCAUSE review* 42.4 (2007), p. 40.

[9] Tatiana A Cardona et al. “Predicting student retention using support vector machines”. In: *Procedia Manufacturing* 39 (2019), pp. 1827–1833.

[10] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”. In: *PeerJ Computer Science* 7 (2021), e623.

[11] Hugh A Chipman, Edward I George, and Robert E McCulloch. “BART: Bayesian additive regression trees”. In: (2010).

[12] Adele Cutler, David Cutler, and John Stevens. “Random Forests”. In: vol. 45. Jan. 2011, pp. 157–176. ISBN: 978-1-4419-9325-0. DOI: 10.1007/978-1-4419-9326-7_5.

[13] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “From data mining to knowledge discovery in databases”. In: *AI magazine* 17.3 (1996), pp. 37–37.

[14] Pratiyush Guleria and Manu Sood. “Data mining in education: A review on the knowledge discovery perspective”. In: *International Journal of Data Mining & Knowledge Management Process* 4.5 (2014), p. 47.

[15] Jennifer Hill, Antonio Linero, and Jared Murray. “Bayesian additive regression trees: A review and look forward”. In: *Annual Review of Statistics and Its Application* 7 (2020), pp. 251–278.

[16] Timothy O Hodson. “Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not”. In: *Geoscientific Model Development* 15.14 (2022), pp. 5481–5487.

[17] Juan Camilo Orduz. *Cohort Retention Analysis with BART - Dr. Juan Camilo Orduz*. Accessed: 1 April 2023. 2023. URL: https://juanitorduz.github.io/retention_bart/.

[18] Osvaldo Martin. *Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ*. Packt Publishing Ltd, 2018.

[19] Osvaldo A Martin, Ravin Kumar, and Junpeng Lao. *Bayesian modeling and computation in python*. CRC Press, 2021.

[20] William M Mason and Nicholas H Wolfinger. “Cohort analysis”. In: (2001).

[21] Martijn Meeter. “Predicting retention in higher education from high-stakes exams or school GPA”. In: *Educational Assessment* 28.1 (2023), pp. 1–10.

[22] Siti Khadijah Mohamad and Zaidatun Tasir. “Educational data mining: A review”. In: *Procedia-Social and Behavioral Sciences* 97 (2013), pp. 320–324.

[23] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[24] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[25] In Jae Myung. "Tutorial on maximum likelihood estimation". In: *Journal of mathematical Psychology* 47.1 (2003), pp. 90–100.

[26] Andy Nguyen, Lesley Gardner, and Don Sheridan. "Data analytics in higher education: An integrated view". In: *Journal of Information Systems Education* 31.1 (2020), p. 61.

[27] Eli Nimy, Moeketsi Mosia, and Colin Chibaya. "Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach". In: *Applied Sciences* 13.6 (2023), p. 3869.

[28] Carlos A Palacios et al. "Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile". In: *Entropy* 23.4 (2021), p. 485.

[29] Posit. *Shiny*. Accessed: 1 October 2023. n.d. URL: https://shiny.posit.co/.

[30] Scikit-learn.org. *3.3. Metrics and scoring: quantifying the quality of predictions*. Accessed: 20 September 2023. 2013. URL: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics.

[31] Mark R Segal. "Machine learning benchmarks and random forest regression". In: (2004).

[32] Dalia Abdulkareem Shafiq et al. "Student retention using educational data mining and predictive analytics: a systematic literature review". In: *IEEE Access* (2022).

[33] SoLAR. *What is Learning Analytics?* Accessed: 10 April 2023. Society for Learning Analytics Research (SoLAR). 2019. URL: https://www.solaresearch.org/about/what-is-learning-analytics/.

[34] Teo Susnjak, Gomathy Suganya Ramaswami, and Anuradha Mathrani. "Learning analytics dashboard: a tool for providing actionable insights to learners". In: *International Journal of Educational Technology in Higher Education* 19.1 (2022), p. 12.

[35] Sandeep Trivedi. "Improving students' retention using machine learning: Impacts and implications". In: *ScienceOpen Preprints* (2022).

[36] Diaa Uliyan et al. “Deep learning model to predict students retention using BLSTM and CRF”. In: *IEEE Access* 9 (2021), pp. 135550–135558.

[37] Seungha Um. *Bayesian Additive Regression Trees for Multivariate Responses*. The Florida State University, 2021.

[38] Elyse Wakelin. “Personal Tutoring in Higher Education: an action research project on how to improve personal tutoring for both staff and students”. In: *Educational Action Research* (2023), pp. 1–16.

[39] Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal. “Mining Education data to predict student's retention: a comparative study”. In: *arXiv preprint arXiv:1203.2987* (2012).

[40] Tianyu Zhang et al. “Application of Bayesian additive regression trees for estimating daily concentrations of PM2. 5 components”. In: *Atmosphere* 11.11 (2020), p. 1233.