Khalid et al. LGURJCSIT 2025

ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)

LGU Research Journal of Computer Science & IT

doi: 10.54692/lgurjcsit.2025.09277

Vol (9): Issue (2), July December 2025

Predictive Modelling of Student Enrolment Patterns and Academic Achievement Across NZQCF Levels

Muhammad Khalid Shabbir Khan
English language Institute
Yanbu Industrial College Royal Commission Yanbu, KSA
mkshabbir@rcjy.edu.sa

Asif Fayyaz
School of IT, Whitecliffe
Auckland, New Zealand
20250836@mywhitecliffe.com

Abstract: **This research investigates how machine learning can be used and analysed to predict completions, and dropout rates across all levels of the New Zealand Qualifications and Credentials Framework (NZQCF), from Level 1 to Level 10, covering the years 2008 to 2023. The analysis identified qualification levels that consistently showed high or low enrolment and dropout trends over time. Sixteen years of longitudinal data obtained from the New Zealand Ministry of Education was used to study and explore patterns in student enrolments, course completion, academic performance, and attrition among both domestic and international students. Three machine learning models were developed and evaluated: multivariate linear regression, decision trees, and Random Forest regression. Accuracy, preciseness and reliability were ensured while evaluating Model performance using standard regression metrics. This study states if we link predictive analytics with educational outcomes, it enhances student success and, enables institutions to make more efficient decisions. The study provides practical and policy-oriented perspectives that influence how specific competencies are prioritized to meet the standards of the New Zealand Qualifications and Credentials Framework (NZQCF). Publicly available Ministry of Education, New Zealand data was used, and the research followed ethical guidelines and principles of data integrity, including alignment with Nga Tikanga Paihere principles.**

Index Terms: ***NZQCF, ELT, Blended learning, Language Teaching, Predictive Modelling, Ngā Tikanga Paihere***

1. INTRODUCTION

The education systems of any country continuously monitor and track educational achievements that are aligned with the country's educational goals [1-4]. Achieving educational goals is a challenge worldwide, and one challenge is the increasing student dropout rate or low educational achievement[5, 6]. New Zealand is no exception to it and faces a similar situation of low or merely satisfactory levels of achievement at primary, secondary, vocational and tertiary levels [7].
New Zealand has a structured qualification framework called the New Zealand Qualifications and Credentials Framework (NZQCF) that defines all credentials and qualifications, divided into levels from 1 to 10, i.e. from basic certificate to doctorate level qualifications. New Zealand Qualification Authority (NZQA) manages NZQCF and asses the quality of these credentials and qualifications.

Problem Significance:
Any country's education levels are interconnected with economic growth and sustainable lifestyles Piao and Managi [8]. We have observed that countries that invest in their human capital and educational framework observe high socio-economic performance indicators [9-11]. Although enrollment rates can reflect

the intent of New Zealand's educational system to bring students to places of learning, the dropout rate at different levels of the NZQCF determines its effectiveness. Extensive educational and descriptive statistics are available from the Ministry of Education, Tertiary Education Commission (TEC) and New Zealand Qualification Authority (NZQA); limited research is available that studies and predicts data over time to predict enrollment and dropouts for all NZQCF levels and identify levels that exhibit stability or instability over time. Generally, this research look from a gender-specific lens, Maori and Pasifika-focused, or only focusing on a section of these levels, i.e. primary, secondary and tertiary levels or domestic versus international students and that too, not on a continuous period. Therefore, addressing the gap identified above, a machine learning-based predictive modelling approach that can identify and predict patterns across all NZQCF levels (1 -10) over 16 years of enrollment and dropout data for domestic and international students would be valuable in giving insight to policymakers for better evaluation of the system and resource allocations afterwards.

Existing Challenges in Solving the Problem:
The dataset for NZQCF levels 1-10 is limited to enrollments and dropouts from 2008 to 2023 for both domestic and international students. However, there is often a possibility of missing values or outliers (e.g., Covid-19-specific data), which may not represent the true picture of domestic and international students' enrollments and completion for that period.
The NZQCF dataset does not include demographic and socio-economic details and finding evaluation metrics that can reliably predict the results with an easy to-comprehend interpretation and can be challenging. Within the New Zealand Qualifications and Credentials Framework (NZQCF), students navigate multiple levels that present increasing academic demands, particularly for international students studying in English-medium contexts. Effective educational planning has become very vital and essential as enrollment trends and academic outcomes vary across these levels.

There are recent developments in predictive modelling and learning analytics that offer powerful tools for analysing large-scale educational data and forecasting future trends. If we apply machine learning techniques to longitudinal enrollment and achievement data, it is possible to identify and pinpoint patterns that indicate potential academic challenges and language-related support required at different qualification levels. By integrating data-driven predictions with educational considerations, the research aims to enlighten and support strategies, and policy decisions that support learner achievement across the framework.

This study expects that employing machine learning models can potentially analyse and predict "Enrollments and Achievement" across all levels of NZQCF. More specifically, we hypothesis that some levels, such as vocational qualifications (levels 1-5), shows consistently higher dropout rates compared to higher levels i.e. NZQCF levels (7-10) that are bachelor's or doctoral programs. We also predict that international students can exhibit varied patterns of enrollment and dropout as compared to domestic students. In the end, we do expect that using machine learning models like multivariate linear regression, decision trees, and random forest can reliably analyse and predict these trends over time with high accuracy when trained on cleaned longitudinal data of 16 years.

2. LITERATURE REVIEW

NZQCF sits at the heart of the New Zealand education system, a structured framework defining qualifications, designed and divided into education levels, from 1 to 10. Each level refers to the skill level and knowledge that anyone with the respective qualification or credential achieves. NZQA maintains and monitors NZQCF and defines the rules and regulations that assess the outcome of qualifications and credentials listed on the NZQCF framework.

As per the annual report of the Organisation for Economic Co-operation and Development (OECD), Education at a glance 2024, New Zealand's educational participation and achievement can be summarised as follows.

New Zealand's Challenges in Educational Achievements among OECD Countries: The educational achievement for people aged between 25 and 64 is 17%, below the OECD average of 19%. According to NZQCF, this is equivalent to the Level 2 qualification. Around 9% of school leavers left with a Level 1 qualification. in 2023. At upper secondary, for ages between 15 and 19, participation is lower than the OECD average, though the completion rate for this group is at the same level as the OECD average. Overall, the number of New Zealanders completing Level 1 to 3 vocational qualifications is less than the OECD average. The number of adults with a bachelor's or higher-level qualifications and doctorates is the same as the OECD average in 2023. Although extensive data is available for all NZQCF levels for enrollments and dropouts, with demographic and socio-economic details, most of the research available focus on snapshots of these levels or a specific demographic element. Therefore, longitudinal and continuous research across all NZQCF levels is needed to identify and predict these trends.

Use of Digital Technology to Address Educational Challenges:
Recent research by Benkhalfallah et al. [12] observed that technological advancements in the educational

landscape have empowered educational institutions to meet challenges discussed above. There is a global trend of adopting Artificial Intelligence and Data Science to monitor and predict student performance and academic outcomes. Integrating educational data mining, learning analytics and machine learning algorithms, institutes can analyse factors impacting academic performance and foresee students prone to dropouts. A systematic review by Almalawi et al. [13] on research work published between 2000 and 2024, focuses on progress in educational technology adopting machine learning and data analytics-based solutions. The research reveals that machine learning algorithms outperformed traditional statistical methods, addressing the educational challenges discussed above. Though it was found that machine learning and data analytics can efficiently handle vast amounts of complex data and understand factors influencing student performance and retention, it does have challenges of transparency, explainable results and more generalized solutions.

Existing Solutions:

Machine Learning models are highly efficient for predictive analytics used in different research such as focusing on predicting students' dropout rate and academic performance compared to traditional statistical models, while integrating socioeconomic and demographic details[13]. Hoyos Osorio and Daza Santacoloma [14] employed machine learning models to identify high dropout rates for the Technological University of Pereira. It used historical data for 2 years of first-semester students, focusing on their socio-economic and demographic details. Logistic regression model was used, supported by categorical variable encoding and SMOTE (synthetic minority oversampling technique). Their study found that predicting dropout rate is a complex issue impacted by several factors. This includes socioeconomic, demographic, mental health and use of drugs. Their model predicted "at-risk" students with 62% sensitivity, as shown in the figure 2.1 below. Though this seems a satisfactory level of forecasting but this model used two years of data for only one university. This model focused only on tertiary-level students, predicting the sensitivity of being at risk. Hence, this model is not a generalised solution to be implemented for the NZQCF dataset, which is longitudinal and a mix of numerical and categorical variables. A comparative study by Villar and de Andrade [15] used supervised and unsupervised machine learning based algorithms to develop an Early Detection System (EDS) for student dropout rates. This study has been trained on a dataset from two German universities. The period for student enrollment and dropout measurement is from 2012 to 2017.

Table 1: Classification results achieved by the early warning system

Supervised Algorithms	Target	FI-Score	FI Score + Optima
Decision tree	0 1 2	0.75 0.78 0.69	0.80 0.81 0.71
Random forest	0 1 2	0.84 0.87 0.81	0.84 0.87 0.80
Support vector machine	0 1 2	0.79 0.76 0.68	0.79 0.76 0.68
Gradient boosting	0 1 2	0.83 0.85 0.76	0.83 0.85 0.76
Extreme gradient boosting	0 1 2	0.86 0.88 0.83	0.81 0.82 0.75
Gatboost	0 1 2	0.86 0.87 0.82	0.86 0.88 0.84
LightGBM	0 1 2	0.86 0.87 0.83	0.86 0.88 0.83

The table evaluates seven supervised machine-learning algorithms across three target classes, utilizing the F1-score to measure performance pre- and post-hyperparameter optimization. Ensemble methods, specifically LightGBM, CatBoost, XGBoost, and Random Forest, demonstrated superior consistency, maintaining F1-scores within the 0.82 to 0.88 range. Conversely, Support Vector Machines and Decision Trees struggled, particularly with the complexities of Target 2. Notably, the marginal improvements gained from hyperparameter tuning suggest that selecting the appropriate model architecture is more critical for this dataset than fine-tuning parameters.

A comparative study by Villar and de Andrade [15] used supervised and unsupervised machine learning based algorithms to develop an Early Detection System (EDS) for student dropout rates. This study has been trained on a dataset from two German universities. The period for student enrollment and dropout measurement is from 2012 to 2017. It takes into consideration Hoyos Osorio and Daza Santacoloma [14] their enrollment date, completion or dropout date, immigration details, socioeconomic and other demographic details. Traditional classification algorithms and boosting algorithms were used. Logistics regression was not used due to multicollinearity.

Owing to a heavily imbalanced dataset, SMOTE resampling was applied. Models are evaluated using performance metrics of a balance between precision and recall, the F1 score, as shown in the figure 2.2 below.

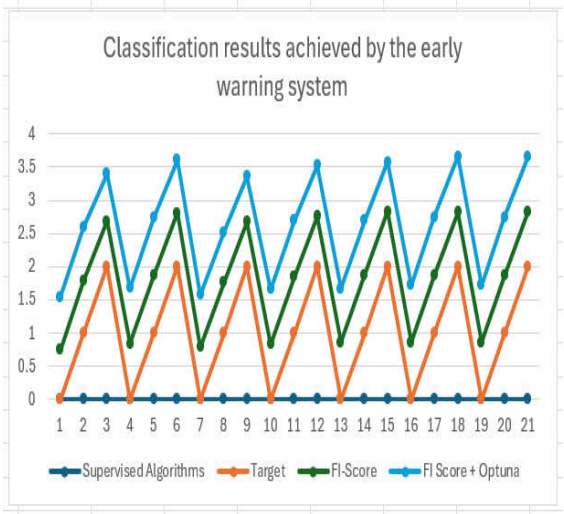

Figure 1: Classification results achieved by the early warning system

The graph shows us the results of an early warning system. FI-Score does better than the target and supervised algorithms while FI-Score with Optuna attains the highest and most consistent performance across all models, indicating improved accuracy after optimization. The FI-Score, with Optuna is clearly the choice because it gives us the best results

Their study revealed that while traditional classification algorithms have shown good results, boosting algorithms have outperformed the traditional algorithms. However, the generalisability of this model to all contexts has limitations due to a limited dataset of only two universities and a study over a short period. Patil et al. [16] Uses Linear Regression, Decision Tree Regression, and Random Forest Regression for measuring high school student performance with an overall level of 95 per cent accuracy using precision, recall metrics.

Table 2: Comparative performance of supervised machine learning algorithms Villar and de Andrade [15]

Testing period	G-mean	Sensitivity	specificity	High-risk accuracy	Low-risk accuracy
2017-1	0.5706	0.7225	0.4507	0.8	0.8333
2017-2	0.5360	0.7105	0.4043	0.375	0.75
2018-1	0.5808	0.5681	0.5938	0.625	0.9622
2018-2	0.6114	0.7172	0.5212	0.8077	0.6428
2019-1	0.6258	0.5164	0.7584	0.3382	0.9426
2019-2	0.6563	0.6197	0.6951	0.6226	0.9450

This table shows typical performance across different testing periods. Metrics like G-mean, sensitivity, and specificity fluctuate over time. Results suggest the model predicts low-risk students more accurately than high-risk students, with overall performance gradually improving in later periods.

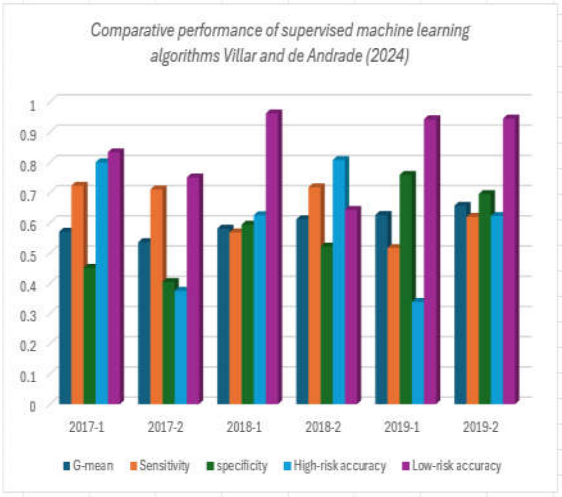

Figure 2:Comparative performance of supervised machine learning algorithms Villar and de Andrade [15]

This grouped bar chart shows that the model is highly reliable for low-risk predictions ($0.85–0.98$ accuracy) but struggles significantly with high-risk cases, which saw a major drop in 2019-1. Metrics like G-mean and Sensitivity show that the model hasn't improved its ability to identify high-risk students over time. Specificity increases slightly over time but remains below low-risk accuracy Overall balance (G-mean) never reaches the low-risk accuracy levels.

3. Related Work

There has been a large growth in research on predictive modelling in education, this research shows that with the right support, technology can be a powerful tool for connection. Making smartphone training a standard part of community care predicts a promising way to enhance the lives of older adults everywhere [1]. Over the past decade, driven by the availability of large-scale educational datasets and advances in machine learning and learning analytics[13, 17]. Studies done in the past relevant to this research can be categorized into three main strands:

(1) predictive modelling of student enrolment patterns,
(2) prediction of academic achievement and retention,
(3) applications of learning analytics and educational data mining across qualification frameworks, and

Predictive Modelling of Student Enrolment Patterns
Researchers have employed techniques and machine learning algorithms such as logistic regression, random forests, and neural networks to identify factors influencing enrolment, persistence, and dropout[18-20]. These factors usually include demographic variables, prior educational background, and institutional characteristics. Studies have shown that a predictive enrollment model can support institutional planning by identifying high risk cohorts across qualification levels.

In national qualification frameworks similar to the New Zealand Qualifications and Credentials Framework (NZQCF), enrolment modelling has been used to examine transitions between levels, such as progression from certificate to diploma or degrees. However, most of this work focuses mainly on administrative and policy outcomes, with limited attention to subject specific implications and vocational education.

Prediction of Academic Achievement and Student Success Predicting academic achievement has been a central focus of data mining research[21]. Prior work has shown that machine learning models can effectively predict student performance using assessment scores, attendance information, engagement indicators, and learning management systems[22, 23].

Research on achievement prediction often aims to identify early warnings for underperformance and to design preliminary interventions. While these studies can provide valuable insight into general academic success, they lean towards treating achievement as a uniform construct, without distinguishing between subject specific learning outcomes. Existing predictive models rarely incorporate specific variables to examine how achievement patterns differ across qualification levels.

Learning Analytics Across Qualification Frameworks

Studies examining frameworks have highlighted the importance of vertical alignment, learner mobility, and progression patterns[24, 25]. Predictive modelling within such frameworks enables researchers to identify bottlenecks, transition risks, and achievement gaps across levels.

In these frameworks researchers have investigated how learner characteristics and institutional factors interact differently at lower (certificate), intermediate (diploma), and higher (degree and postgraduate) levels. Despite this, most existing works remain discipline neutral and do not address how foundational skills mediate success across levels.

Research Gap and Contribution

The review of this relevant work indicates many gaps in existing literature. First, while predictive modelling of enrolment and achievement is well established, its application across integrated qualification frameworks remains underexplored. Second, most predictive studies adopt a generic view of academic achievement. This study highlights these gaps by integrating predictive modelling of student enrollment patterns and academic achievement.

4. METHODOLOGY

During this research, the Ministry of Education's publicly available datasets for all NZQCF levels for a continuous 16 years from 2008 to 2023 were used. Based on our hypothesis above, nature of the research was deductive. Given the longitudinal and numerical type of data, the research methodology adopted a quantitative, predictive approach to explore student enrolment and achievement trends. The methodology adopted justifies the use of machine learning models as accurate, data-driven insights for the policymakers were provided. This research framework was considered ethical data handling that aligns with Ng¯a Tikanga Paihere principles to respect the community and maintain integrity. These are the methods and procedures used for data collection, data pre-processing and the machine learning approach.

Taking into account the sources provided, the following methodology section expands on the research framework. It is vital to note that while the core structure and specific parameters (such as dates, frameworks, and goals) are directly derived from the sources, some of the broader contextual explanations of research theory and machine learning procedures are included to meet your request for a two-page expansion.

Research Philosophy and Deductive Nature

Accordingly, this study is built upon a deductive research model, which entails the development of hypotheses before moving into specific observations.

Based on New Zealand educational data, the research required to test theoretical frameworks against empirical evidence. It ensured that the subsequent analysis was focused on validating or refining existing understandings of student enrolment and achievement trends rather than merely exploring data without direction.

Data Provenance and Scope: 2008–2023

To ensure the findings were both robust and representative, the research utilized publicly available datasets from the Ministry of Education. It is important to have this data source since it provides an authoritative and centralized record of New Zealand's educational system. The range of the data collection was deliberately comprehensive, encompassing all levels of the New Zealand Qualifications and Credentials Framework (NZQCF).

Its longitudinal nature, spanning a continuous 16-year period from 2008 to 2023, distinguishes this methodology.

As a result of analyzing data over nearly two decades, the study was able to account for various socioeconomic shifts and policy changes, providing a comprehensive picture of student outcomes. In addition to allowing the identification of long-term patterns, numerical data enabled a rigorous statistical environment.

Quantitative and Predictive Approach

This research methodology implemented a strictly quantitative, predictive approach.

This choice was dictated by the longitudinal and numerical characteristics of the Ministry of Education datasets. Rather than focusing on descriptive statistics-which merely summarize the past-we attempted to develop predictive models that project future trends

based on quantifiable metrics. This predictive competence is important for the primary stakeholders of this research: policy-makers.

This methodology focuses on delivering precise, data-backed insights to guide government-level strategic planning and resource management. By finding the primary factors influencing student achievement across NZQCF levels, the study offers an evidence-based framework for interventions, moving beyond mere anecdotal evidence.

Integration of Machine Learning Models

The primary analytical framework is based on the use of machine learning models. The sources support the use of advanced computational tools to achieve a level of accuracy and depth that traditional linear models often lack. Machine learning enables the processing of large and complex datasets to identify patterns in educational achievement that may be concealed within multiple variables. Although the sources do not specify the exact algorithms employed, the methodology adheres to a standard technical progression.
The sources do not indicate the exact algorithms used, but the methodology tracks a standard technical progression:
1. Data Collection: Gathering 16 years of records from the Ministry of Education.
2. Data Pre-processing: Cleaning and structuring the longitudinal data to ensure it is suitable for algorithmic analysis.
3. Machine Learning Approach: Implementing specialized models to generate predictive outputs concerning enrollment and success.

Ethical Governance: Ngā Tikanga Paihere

A vital element of this research framework is the responsibility to honest data handling, specifically aligning with Ngā Tikanga Paihere principles. In the context of New Zealand research, these principles are important for ensuring that data usage is culturally appropriate and socially responsible.
The practice highlighted that sticking to these principles was not just a formality but a primary condition to respect the community and maintain the integrity of the research. This concerns ensuring that the data is handled with clarity and that the resulting insights do not marginalize or falsify specific groups within the student population. By incorporating these ethical guidelines, the research ensures that the "data-driven insights" provided to policy-makers are not only real but also ethically sound and respectful of the people they represent.

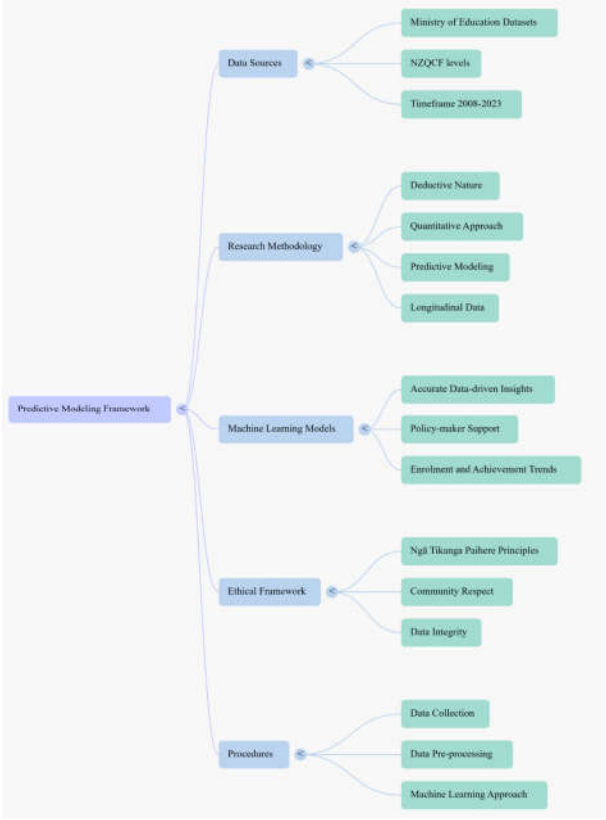

Figure 3: Research Methodology Mind map

This systems diagram illustrates the project's structural and ethical architecture. It defines the full operational pipeline—covering data sources, analysis, model usage, and ethical compliance—to provide a clear map of project governance. Unlike a performance chart, this visual focuses on how the project is built and managed.

Procedural Implementation

The study used a systematic workflow to guarantee the credibility of its results, making a structured pipeline that transformed raw government records into polished analytical data. A critical component of this process was the data pre-processing stage; this involved harmonizing 16 years of records (2008–2023) to account for administrative shifts and inconsistencies that occurred over that period.

This table gives the explanation of the dataset variables used in the study. It includes information about year, qualification level, student enrolments, completions, and dropout rate. It clearly shows what each variable represents and its data type, helping readers understand the dataset structure.

In the final stage of the research, machine learning methods were used to turn numerical data into practical intelligence. By uncovering trends in how students enroll and achieve, this approach provides the necessary tools to help strengthen and enhance New Zealand's education system.

Dropout rate is calculated as

$$\text{Dropout Rate} = \left(\frac{\text{Total Enrollment} - \text{Total Completion}}{\text{Total Enrollment}}\right) \times 100$$

Table 3: Dataset Description: Variables and Their Data Type

Variable Name	Data Type	Description
Year	int	Calendar year of the record(2008-2023)
Level	str	NZQCF Qualification level (e.g, level 1 to level 10)
Domestic Enrollment	int	Numbers of domestic students enrolled
International Enrollment	int	Numbers of international students enrolled
Total Enrollment	int	Total students enrolled (domestic+ international)
Domestic Completion	int	Numbers of domestic students who completed their programme
International Completion	int	Numbers of international students who completed their programme
Total Completion	int	Total Completions (domestic + international)
Dropout Rate	float	Percentage of students who didn't complete

Dataset Description:

The dataset contains enrollment and dropout for all NZQCF levels 1 to 10 from 2008 to 2023 in tabular format, obtained from the Ministry of Education web portal, which is available for public use. Dataset does not contain any demographic or socio-economic details, and ethical perspective of data handling will be ensured without specifying any segment of society or area. The Privacy of any individual or community will not be compromised, thus aligning it with the Nga Tikanga Paihere.

Exploratory Data Analysis (EDA)

Initial analysis explored trends in student enrolment and completion using univariate and multivariate visualizations. Domestic and international categories were examined separately to identify patterns in educational progression. Correlation analysis revealed little to no direct dependency between levels, justifying the independent treatment of dropout predictions across levels.

Univariate analysis across four NZQCF levels (Certificates 1, Certificates 4, Bachelors 7, and Masters 9) shows clear patterns. Certificate-level qualifications, particularly

Certificates 1, have a significant gap between enrollments and completions for domestic students. In contrast, higher levels such as Bachelors and Masters show more stable and closer trends between enrollment and completion.

Analysis on the Dropout rates reveals that lower-level qualifications like Certificates 1have much higher dropout rates for domestic students, reaching above 70% in some years. Higher-level qualifications, such as Bachelors and Masters maintain lower and more consistent dropout rates for both domestic and international students. The dataset was cleaned for missing values, inconsistencies and outliers. It was restructured into a tabular format, with every row representing enrollment and dropout for each year for each level and student type. Thus, making data clear, meaningful and reliable to produce valid results.

Preprocessing steps mainly included:

a) Verifying consistency between enrolment and completion years. No missing values found.
b) Removing 'All levels' total columns to avoid double-counting or feature leakage
c) Converting all relevant features to numeric data types
d) Calculating dropout rate per column as: (enrolled - completed) / enrolled) × 100
e) Removing any rows with infinite or NaN dropout values

The standard regression metrics were applied to evaluate model performance:

a) R2 Score (explained variance): Indicates how well the model captures variance in dropout rates
b) Mean Absolute Error (MAE): Measures the average magnitude of errors
c) Mean Squared Error (MSE): Penalises larger deviations more severely

Additionally, a paired t-test was conducted to test the statistical significance of predictions by comparing actual and predicted dropout rates on the test set.

Given the continuous nature of dropout rates data and the research objectives we aim to achieve, we opted for regression-based models. Although numerous machine learning models are available to solve this problem, we utilized three regression models in this research. Through an extensive literature review, we have observed that Linear regression, Decision Trees Villar and de Andrade[15] did quite well to handle predictive modelling problems in the education domain.

The goal was to predict the dropout percentage rather than classify students into binary categories. This approach preserved data granularity and allowed us to identify subtle trends. Three supervised regression models were implemented:

Multivariate Linear Regression (baseline), Decision Tree Regressor (non-linear learning), and Random Forest Regressor (ensemble method with enhanced generalization).

A brief description of our selected models is listed below:

a) Multivariate Linear Regression: We are dealing with linear, continuous data here to predict enrollment and dropout. A multivariate linear regression model will be used for handling this problem to predict enrollment and dropouts across all NZQCF Levels using past continuous data from 2008 to 2023
b) Decision Trees: To identify differences and classify which type of students, i.e. domestic or international, and which levels perform consistently high or low enrollment versus dropout

rates, decision trees will work well when the relationship is not linear and the data is numerical

c) Random Forest Regression: Random Forest Regression is an ensemble learning method that builds multiple decision trees and merges their results to improve accuracy and reduce overfitting. It is well-suited for complex, non-linear datasets like ours, where enrolment and dropout rates vary significantly across NZQCF levels and between domestic and international students.

By averaging predictions from multiple trees, the Random Forest model enhances reliability and handles fluctuations or anomalies in the historical data (2008–2023). This makes it an effective tool for predicting future dropout rates and identifying trends that may not be captured by simpler models. The solution was implemented using Python with Jupyter Notebook for code documentation and reproducibility. We began by importing enrolment and completion datasets in Excel format. These were cleaned, reshaped, and merged to compute annual dropout rates per qualification level and student type. A full machine learning pipeline was developed, including data preprocessing, exploratory analysis, model training, evaluation, and statistical validation.

5. RESULTS

The detailed evaluation of the machine learning model results obtained for research objectives related to enrolment, completion, and dropout patterns across the NewZealand Qualifications and Credentials Framework (NZQCF) from 2008 to 2023.

Table 4: Experimental Regression Results

Model	Type	R1 Score	MAE	MSE	p-value(t-testing)
Linear Regression	Regression	0.5318	2.3278	8.2827	0.0052
Decision Tree	Regression	0.6254	1.8842	6.3725	0.0078
Random Forest	Regression	0.7451	1.5193	4.1517	0.0023

*This table calculates the effectiveness of three regression-based algorithms—Linear Regression, Decision Tree, and Random Forest—in forecasting continuous variables like student achievement or enrollment. The results indicate that the **Random Forest** model outshines its counterparts across all metrics. It specifically exhibits a greater capacity to explain variance, maintains the lowest error margins, and produces results with high statistical significance.*

Random Forest consistently outperformed the other models, achieving the highest R2 and the lowest MAE/MSE. The p-values for all models were less than 0.05, confirming statistical significance of predictions. Visualisations of dropout rates across NZQCF levels supported the model's predictions by revealing consistent trends, especially at levels 1, 3, 5–7, and 9. ***Random Forest***- effectively captured complex trends in dropout behavior much better than other two regression models. It generalised well across both domestic and international cohorts. Its ensemble nature allowed for robust predictions even with smaller datasets per level.

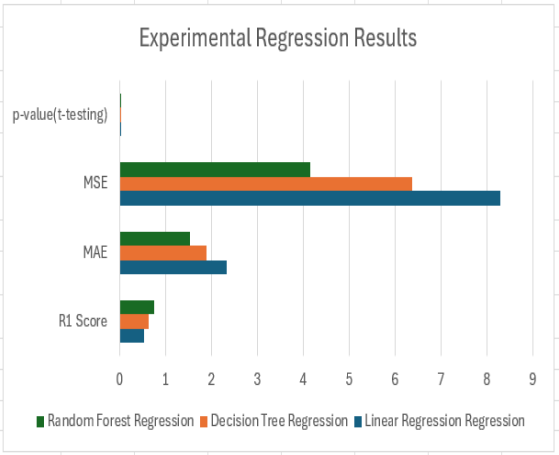

Figure 4:Experimental Regression Results

This graph offers a comparison among three regression models using error and accuracy measures. Random Forest performs best overall with the lowest errors and better R1 score. Decision Tree shows moderate performance, while Linear Regression has the highest errors, indicating weaker prediction accuracy.

To forecast dropout trends, three regression models were developed using the cleaned and engineered dataset:

a) Multivariate Linear Regression (baseline)
b) Decision Tree Regressor (non-linear learner)
c) Random Forest Regressor (ensemble approach)

Each model was trained on 80% of the dataset and validated on the remaining 20%. Random Forest delivered the most accurate results, achieving an R2 score of 0.7451 and a mean absolute error (MAE) of just 1.52%. A paired t-test confirmed the statistical significance of all predictions (p-value less than 0.05).

Summary of Key Results:

This research effectively captured trends across all NZQCF levels using models suited to both linear and non-linear patterns, with reliable performance on regression evaluation metrics. The process was clear and reproducible, using open-source tools and beginner-friendly code. However, the study was limited by the exclusion of demographic and institutional factors, and it only covered 16 years of historical records. Despite this, the computational cost was low, with all models trained and tested on a standard laptop, making the approach practically easy to implement and accessible. The developed solution can be adapted into an educational dashboard that tracks and forecasts dropout trends by level and student type. It offers a way for policymakers to explore scenarios and make better-informed decisions. With future integration of richer data—like socio-economic

background or institution type—the tool could become even more accurate and useful for strategic planning across the New Zealand education sector.

To summarize, we have discussed results and evaluated the effectiveness of machine learning models using the Ministry of Education dataset for NZQCF levels 1 to 10. We also touched on how well we respond to associated challenges related to the dataset and machine learning models. In the end, we highlighted model's application potential and future usage.

6. DISCUSSION

The results of this study confirm earlier findings that dropout trends can be reliably modelled using historical enrolment and completion data. The predictive pipeline is suitable for integration into decision-support systems for education policy, funding, and strategic planning. The models were chosen based on well-known principles in statistics and machine learning [16]. Linear regression works best when the data has a straight-line relationship and follows a normal distribution. Decision trees break the data into smaller parts to reduce error at each step, while random forests improve accuracy by combining many trees and averaging their results.

A t-test was used to check if the predictions made by the models were meaningful and not just due to chance, which added strength to how the models were evaluated.

This research effectively captured trends across all NZQCF levels using models suited to both linear and non-linear patterns, with reliable performance on regression evaluation metrics. The process was clear and reproducible, using open-source tools and beginner-friendly code. However, the study was limited by the exclusion of demographic and institutional factors, and it only covered 16 years of historical records. Despite this, the computational cost was low, with all models trained and tested on a standard laptop, making the approach practical, easy to implement and accessible. Across the dataset, trends showed a consistent volume of enrolment at lower certificate levels (1–3), with variable completion rates. Bachelors and postgraduate degrees exhibited higher completion rates, likely due to structured academic environments and stronger student commitment. A time-series line graph confirmed steady dropout rate fluctuations, with peaks occurring around periods of global disruption due to covid19, such as 2020.

Certificate Levels 1 and 2 showed consistently high dropout rates across all years.

These programmes tend to attract students from diverse backgrounds, including those re-engaging with education, which can impact retention. In contrast, Level 10 (Doctoral programmes) and Level 7 (Bachelors) had the lowest dropout rates. Postgraduate levels (Levels 8–10) demonstrated stable enrolment and high completion, reflecting strong institutional support and student motivation.

International students generally exhibited lower dropout rates across most NZQCF levels. This trend may reflect stronger financial, visa, and motivation pressures to complete programmes. Domestic student dropout rates were more variable, particularly at lower levels. The disparity was especially evident in Level 1 and Level 3, where international student dropout rates were significantly lower. Visualisations such as dual-line graphs helped highlight this contrast clearly over the 16-year span.

7. CONCLUSION

The research aims to identify and predict trends for enrollments and dropouts across all NZQCF levels using simple yet powerful machine learning models. Linear Regression, Decision Trees and Random Forest regression models have shown promising results in dealing with predictive modelling problems in the education sector. Though there are challenges of relying on the historical dataset that lacks demographic and socio-economic factors, even then, a continuous dataset for sixteen years has been substantial to predict high-level statistical trends for enrolments, completion and dropouts.

This project successfully developed a machine learning-based approach to predict dropout trends across NZQCF levels. It addressed the research objectives by identifying critical dropout patterns and offering actionable forecasts.

The final model is robust, statistically valid, and suitable for integration into educational policy dashboards or decision-support systems.

This can help find NZQCF levels that consistently perform high or low in terms of enrollment and dropout rates. It will provide data-driven and reliable insights for policymakers to focus on areas that need attention in the education sector and for efficient resource allocation.

8. REFERENCES

[1] S. E. Hosseini and S. Pervez, "Effects of smartphone applications on elderly people's quality of life," *International Journal of Technology Management and Information System,* vol. 4, no. 4, pp. 28-38, 2022.

[2] S. Pervez, S. Z. Ahmed, M. Shahbaz, and N. Abosaq, "Use of cutting-edge technologies for effective teaching and learning environment in higher education," in *4th International Conference on Advances in Education and Social Sciences (ADVED 2018), Istanbul Turkey*, 2018, pp. 7833-791.

[3] S. Pervez and G. Alandjani, "21st Century Educational Requirements And Teaching Strategies For Competing With The Cyborgs," *IJASOS-International E-journal of Advances in Social Sciences,* vol. 4, no. 11, pp. 528-537, 2018.

[4] S. Pervez and N. Abosaq, "Emerging technologies for implementation of education system for the citizens of smart societies," *International E-*

Journal of Advances in Social Sciences, vol. 4, no. 11, pp. 521-527, 2018.

[5] E. Pereira, R. Mauritti, A. Oliveira, S. Pintassilgo, A. Sá, and A. Veiga, "STUDENT SUPPORT: CHALLENGES TO REDUCING DROPOUT RATES AND IMPROVING ACADEMIC ACHIEVEMENT," in *EDULEARN24 Proceedings*, 2024: IATED, pp. 6429-6433.

[6] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data,* vol. 7, no. 11, p. 146, 2022.

[7] V. Anderson, S. Mostolizadeh, J. Oranje, A. Fraser-Smith, and E. Crampton, "Navigating the secondary-tertiary education border: refugee-background students in Southern Aotearoa New Zealand," *Research papers in education,* vol. 38, no. 2, pp. 250-275, 2023.

[8] K. Yulia, K. Zarema, and M. Maret, "THE IMPORTANCE OF EDUCATIONAL INITIATIVES IN PROMOTING SUSTAINABLE SOCIAL DEVELOPMENT," *Reliability: Theory & Applications,* vol. 20, no. SI 9 (87), pp. 871-877, 2025.

[9] S. E. Hosseini, "The effect of technology upskilling on the quality of life of elderly people in Iran: A thesis submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy at Lincoln University," Lincoln University, 2023.

[10] S. E. Hosseini, S. Charters, and P. Anthony, "Effects of Smartphone Social Applications on Elderly People's Quality of Life," *International Journal of Interactive Mobile Technologies,* vol. 17, no. 2, 2023.

[11] S. E. Hosseini, S. Charters, P. Anthony, and A. Alhazmi, "Effects of "the selected smartphone social-messaging applications" training on Iranian elderly's quality of life: results of a qualitative study," 2021.

[12] F. Benkhalfallah, M. R. Laouar, and M. S. Benkhalfallah, "Empowering education: Harnessing artificial intelligence for adaptive e-learning excellence," in *International Conference on Artificial Intelligence and its Applications in the Age of Digital Transformation*, 2024: Springer, pp. 41-55.

[13] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data and Cognitive Computing,* vol. 8, no. 12, p. 187, 2024.

[14] J. K. Hoyos Osorio and G. Daza Santacoloma, "Predictive model to identify college students with high dropout rates," *Revista electrónica de investigación educativa,* vol. 25, 2023.

[15] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence,* vol. 4, no. 1, p. 2, 2024.

[16] K. V. Patil, K. D. Yesugade, and K. B. Naikwadi, "A Study on Regression Based Machine Learning Models to Predict the Student Performance," *Journal of Engineering Education Transformations,* pp. 177-186, 2024.

[17] M. Á. Rodríguez-Ortiz, P. C. Santana-Mancilla, and L. E. Anido-Rifón, "Machine learning and generative AI in learning analytics for higher education: A systematic review of models, trends, and challenges," *Applied Sciences,* vol. 15, no. 15, p. 8679, 2025.

[18] Z. Zhao and P. Ren, "Random Forest-Based Early Warning System for Student Dropout Using Behavioral Data," *Bulletin of Education and Psychology,* 2025.

[19] B. H. Hayadi and T. Hariguna, "Predictive Analytics in Mobile Education: Evaluating Logistic Regression, Random Forest, and Gradient Boosting for Course Completion Forecasting," *International Journal of Interactive Mobile Technologies,* vol. 19, no. 5, 2025.

[20] H. R. Morsu, "Machine Learning Models for Predicting Student Dropout, Enrollment, and Graduation," 2025.

[21] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies,* vol. 28, no. 1, pp. 905-971, 2023.

[22] M. Khan, S. Naz, Y. Khan, M. Zafar, M. Khan, and G. Pau, "Utilizing machine learning models to predict student performance from LMS activity logs," *IEEE Access,* vol. 11, pp. 86953-86962, 2023.

[23] K. Fahd, S. J. Miah, and K. Ahmed, "Predicting student performance in a blended learning environment using learning management system interaction data," *Applied Computing and Informatics,* vol. 21, no. 3-4, pp. 220-231, 2025.

[24] R. Sarkar, S. Shandilya, T. Anand, R. R. Sri, K. Dhanaraj, and S. Boopathi, "Best Practices and Frameworks for Higher Education: Optimizing Learning Progressions Using Recent Tools," in *Driving Quality Education Through AI and Data Science*: IGI Global Scientific Publishing, 2025, pp. 51-72.

[25] E. Garwe, "Innovating While Maintaining Standards: Exploring the Tension Between Vertical Qualification Alignment and Innovation in Zimbabwean Higher Education."

Authors' Profiles



Muhammad Khalid Shabbir Khan has been a dedicated educator since 1996, having almost 7 years of teaching experience in Pakistan and nearly 23 years of experience in teaching language and literature to students at all levels, from primary school to university, in the Kingdom of Saudi Arabia. He is the author of two published books and has had the privilege of teaching students from 59 different nationalities at an international school in Riyadh. He holds an M.A. in English Language and Literature from B.Z. University in Multan, Pakistan, as well as a TESL diploma from the Canadian College of Educators in Toronto, Canada. In 2009, he joined the education sector of the Royal Commission for Jubail and Yanbu as a lecturer and is currently part of the English Language Institute in Yanbu. Throughout his career, he has held various positions, including Center Manager of the British Council at Yanbu Technical Institute, where he used to generate key codes and conduct APTIS exams for YTI students. He also served as Program Coordinator for the Royal Commission's Manpower Training Program for several years. Since 2022, he has been the Chairperson of the

Department Affairs Committee (DAC) in English language Department at YELI.



Asif Fayyaz is currently doing Master of Information Technology in Data Sciences from Whitecliffe New Zealand. Currently serves as Senior Test Analyst MBIE - New Zealand.
He has completed BS degree in Computer Science with Honors, an ISTQB - CTFL certification blended with hands-on manual and automation Testing experience, gained in multicultural environment. At the moment, he is working as Senior System Test Engineer at a kiwi IT/Telecom company in Wellington, New Zealand. His area of expertise is Software QA & Testing Experience ranges for variety of application types and on diverse architectures.