Predictive Regression Modeling for Forecasting Graduation Duration in Online Offsite Degree Program

Buddhini Gunarathna, Vishaka Nanayakkara, Buddhika Karunarathna and Tharanee De Silva
University of Moratuwa, Colombo, Sri Lanka

bjgunaratna1996@gmail.com
vishaka@cse.mrt.ac.lk
buddhika@cse.mrt.ac.lk
tharaneesarangika@gmail.com

Abstract: The demand for Information Technology (IT) professionals continues to rise across various sectors, where they play vital roles. However, the supply of IT graduates often fails to meet industry needs and this is a huge problem for the Sri Lankan IT Industry (National IT-BPM Workforce Survey – 2019). In this context, this study presents a predictive regression modelling approach to predict graduation duration in the Bachelor of Information Technology (BIT) degree program at the University of Moratuwa, Sri Lanka. It integrates demographic data—student district, birth year, AL results, OL maths grade, gender, employability status, occupation, and AL stream—along with academic performance indicators like diploma completions and higher diploma completions. After evaluating the suggested features, the key findings indicate the significance of certain features, notably the number of semesters taken to complete the diploma, higher diploma, and the degree. Additionally, demographic factors such as district, birth year, AL results, OL maths grade, gender, and employability status were found to be important. The regression analysis was carried out using the Orange data mining tool (Orange Data Mining). Various algorithms, including random forest, neural network, linear regression, and k-nearest neighbours (kNN), were used to develop predictive models. By adjusting parameters such as metrics, weights, number of neighbours, number of iterations, and training dataset size, the models were optimised to better fit the dataset. Training and testing the models revealed consistent error metrics, including MSE, RMSE, MAE, and R^2, validating the accuracy of predictions. By considering the least and reasonable error in each model, the most suitable model to fit the given dataset was selected. The prediction model accurately forecasted graduation duration for subsequent academic batches, demonstrating its effectiveness in predicting student progress in the program. This research contributes to understanding the factors influencing graduation duration in a distance learning context and provides insights for educational institutions to optimise program planning and student support initiatives. Additionally, it is a good indicator to the companies to gain a better understanding of the availability of future workforce.

Keywords: Predictive analysis, Regression model, Graduation duration, Future workforce, IT industry

1. Introduction

The Bachelor of Information Technology (BIT) degree program is offered by the Center for Open and Distance Learning (CODL) at the University of Moratuwa. The BIT is a three-year degree program introduced primarily for students who have passed GCE Advanced Level (AL) examination in Sri Lanka. This is conducted in the open and distance learning mode. Students use Moodle (Moodle.org) as a learning management system, in a way that students can log in and access uploaded lecture materials. Lecture recordings, lecture notes, and both formative and summative assignments are provided for the students. Formative assignments do not count towards the final grade but are intended to enhance students' understanding. Summative assignments contribute to the final grade. At the end of each semester, students must achieve an average of over 40 from the summative assignments to be eligible to sit for the final examination. The final grade is calculated based on the total of the summative assignments and the final examination.

The programme consists of two semesters per year, resulting in a total of six semesters for the entire program. The programme has two intakes per year for the degree program. The batches are named as follows:

Table 1: The Naming of Student Batches

The batch name	Started Year
18S1	23/04/2018 (in Early 2018)
18S2	22/10/2018 (in mid-2018)
19S1	27/05/2019 (in mid-2019)
19S2	30/12/2019 (in end-2019)

A new curriculum was introduced in 2018, featuring 11 compulsory modules for each of the first two years and 15 modules for the third year. The third year comprises four compulsory and 11 elective modules. If students successfully complete all the modules in the 1st year with an average GPA of over 2.0, they qualify for the Diploma Certificate in IT by the end of the first year. The effective date of the diploma is determined after the completion of the last exam in the 1st year. Similarly, if students complete all modules in both the 1st and 2nd years with an average GPA of over 2.0, they are eligible for the Higher Diploma certificate in IT upon completion of the 2nd year. The effective date of the Higher Diploma is decided based on the completion date of the last exam in the 1st and 2nd years. Furthermore, to obtain the Bachelor of Information Technology degree, students are required to complete all the compulsory modules and, six electives in the third year attaining a semester grade point average (SGPA) above 2.0.

Additionally, students have a total of six years, equivalent to 12 semesters, from the date of registration to complete the program and an additional three years are provided for program completion. Failure to meet this extended deadline may result in an inability to graduate. Based on this, the final academic semester for the 18S1 batch to complete the degree is 23S2. Only students from the 18s1 to 19s1 batches who faced the new curriculum, had the opportunity to graduate by January 2023 as they had completed 6 semesters.

On the other hand, the Information Technology field is rapidly evolving, with sectors like healthcare (Evolving technology in health sector), academia (Technology in Academia: Trends Shaping the Future of Higher Education), and others increasingly dependent on information technology solutions. Transitions from manual to automated systems happen rapidly and so the demand for IT professionals grows. Many people with an IT background are involved in organisations and institutes as Data Analysts, Business Administrators, System Analysts, Data Entry Operators, and so on. Our main goal in the BIT degree program at the University of Moratuwa is to produce IT graduates for the industry. However, the supply of IT graduates remains insufficient to meet industry demands. According to the 2019 survey of National IT BPM workforce (National IT-BPM Workforce Survey - 2019) conducted by the Information and Communication Technology Agency of Sri Lanka, there is a gap of nearly around 12,000 between demand and supply in Sri Lanka. Predicting the number of IT graduates is crucial for staff planning, work planning. This research aims to develop a predictive regression model to forecast the number of semesters students will take to complete their degree, thereby predicting the number of future graduates who will complete the degree program. The belief is that this method can be used for other degree programs as well to predict the number of semesters students will take to complete their degree and forecast the future workforce.

Various studies have predicted graduation rates, student behaviors, and dropout rates at universities. Goenner et al. conducted a predictive analysis (Predicting graduation rates at Doctoral Universities) to understand the factors affecting graduation rates at 258 doctoral universities. They considered factors such as student demographics (age, state, scores) and institutional characteristics (percentage of PhD holders, educational expenditures, student-faculty ratio). In their study they suggest to enhance graduation rates by improving the identified important factors. According to Lidya et al. (Predicting student graduation - 2024), predictive models benefit both students and institutions by facilitating better decision-making. Students can make informed choices about their career paths, while institutions can identify students at risk of not graduating and take proactive measures to improve institutional performance.

The research aims to predict graduation duration using various models and identify the most effective model for prediction.

2. Objectives

The primary objective of this research paper is to predict the number of semesters needed for students to complete the Bachelor of Information Technology (BIT) degree program, drawing data from cohorts enrolled in the 18S1, 18S2, and 19S1 batches. Utilising regression analysis, the study will employ various models to assess their efficacy and select the most suitable model for the dataset. Subsequently, the chosen model will be used to forecast the completion time for future BIT students. Furthermore, leveraging this forecast, the research aims to estimate the number of graduates from the BIT program to facilitate strategic planning and ensure alignment with industry needs. Through these objectives, the study seeks to provide valuable insights into academic progression and workforce preparedness in the field of information technology.

3. Methodology

3.1 Tools

Orange Data Mining software (Orange Data Mining) is utilised in this project to enhance and support the objectives. Orange is a free and open-source software that features a user-friendly workspace for drag-and-drop functionality, allowing users to modify parameters easily. The software offers various tools, including file import, save, view data table, and Python script tools. It also comes with pre-built machine learning models, simplifying the process for users allowing them to adjust the parameters according to the specific dataset. This enables users to train the dataset with various models to enhance efficiency and select the most appropriate one. Additionally, the software enables forecasting of important features based on the designed diagram.

3.2 Feature Selection

The target here is the number of semesters taken by a student to complete the degree. For that, selecting appropriate features for a model is crucial, as the weighting of these features significantly influences the model's construction. Initially, a set of features for our dataset is identified to facilitate model fitting. The primary focus was on students' performance indicators, including the effective dates of diploma and higher diploma attainment. Using these effective dates, the number of semesters taken to reach each educational milestone was calculated from the registration date. Additionally, demographic data such as district of residence, birth year, Advanced Level (AL) examination results, AL stream, z-score, Ordinary Level (OL) mathematics grade, gender, and employability status were taken into account. Prior to incorporating these features into the model, a preliminary test is conducted to evaluate their impact using the Orange software and generated diagrams to examine the variation of the target variable in relation to these features.

		#	Univar. reg.	RReliefF
1	District	23	NA	0.693
2	OL Maths	4	NA	0.486
3	Z Score		NA	0.385
4	S		NA	0.351
5	Gender	2	NA	0.351
6	H.D		NA	0.298
7	Diploma achieved semester		NA	0.293
8	C		NA	0.276
9	Employed	2	NA	0.177
10	B		NA	0.154
11	A		NA	0.060
12	Birth Year		NA	0.046

Gender
Categorical with 2 values: 'FeMale', 'Male'

Figure 1: The impact on the dataset occurs due to the selected features

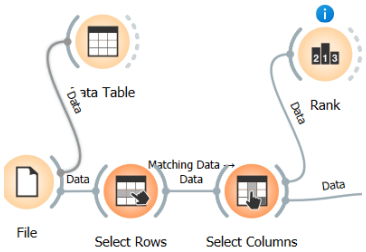

Figure 2: Feature selection

The RReliefF algorithm is utilised as the feature selection algorithm using the Orange data mining tool. The RReliefF algorithm considers both redundancy and relevance, producing a score based on these factors. In the dataset, every feature is attributed a score, and the Rank tool discerns the most pertinent features for the dataset based on this score, analysing the interaction between each feature and the target variable. This approach is efficient for feature selection in a given dataset. Figure 1 illustrates that the District of Residence emerged as the most influential feature in our dataset, exhibiting a high RReliefF score. This result may be attributed to the higher concentration of students from the Colombo District, where the Moratuwa University offers the BIT external degree. The accessibility of Moratuwa University to students residing in the Colombo

District, coupled with the reputation of the BIT external degree program, likely contributes to this observed impact.

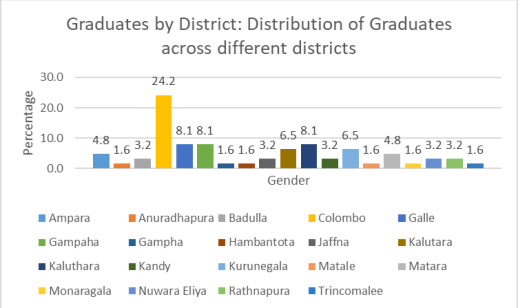

Figure 3: Graduates by District: Distribution of Graduates across different districts

Figure 3 illustrates that the highest proportion of graduates, 24.2%, are from the Colombo district. The second-highest percentage is from the districts of Galle, Gampaha, and Kalutara, at 8.1%, indicating a significant disparity. This suggests that the district of residence directly influences the final number of graduates.

Figure 1 shows that the Ordinary Level (OL) Mathematics grade and gender also significantly influenced the model. A majority of students who received an 'A' grade in OL Mathematics successfully completed the degree, with the majority being male.

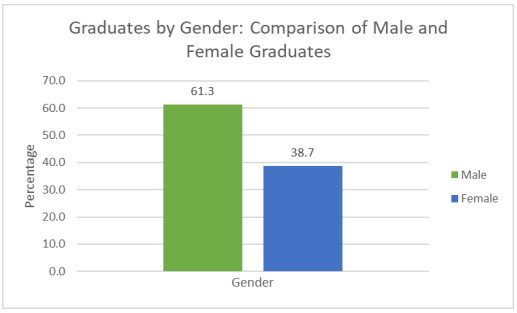

Figure 4: Graduates by Gender: Comparison of Male and Female Graduates

In Figure 4, the majority of graduates are male, comprising 61.3% of the total. Females account for 38.7% of the graduates in the BIT degree program. There was no initial differentiation based on gender at the time of registration; however, this disparity is observed in the final graduation rates.

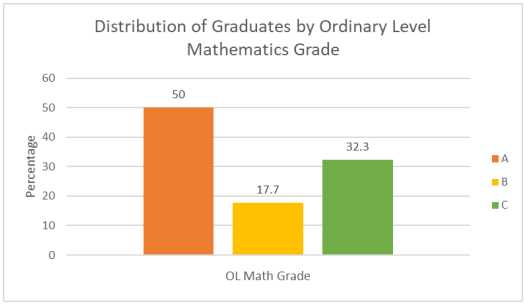

Figure 5: Distribution of Graduates by Ordinary Level Mathematics Grade

The grade in Ordinary Level (OL) Mathematics is considered a measure of students' mathematical aptitude at the time of registration for the BIT degree program. As a primary prerequisite, students are required to achieve a grade above C in their OL mathematics exam to qualify for enrollment in the BIT degree program. Figure 5 illustrates that 50% of the graduates attained an A grade in mathematics. Those who received a B grade accounted for 17.7% of the total graduates, while 32.3% had obtained a C grade and successfully completed the BIT degree program.

The z-score serves as a metric to assess intelligence among students who pursued the Sri Lankan Advanced Level (AL) curriculum across different streams in Sri Lanka. The Rank tool indicate the significance of the z-score as a feature, assigning it a RReliefF score of 0.385 and identifying it as the third most important feature.

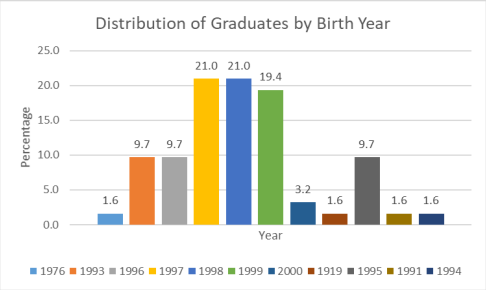

Figure 6: Distribution of Graduates by Birth Year

Figure 6 presents a simple analysis indicating that the highest number of students who graduated from the BIT degree program between 2018 and 2023 were born in 1997 and 1998, each accounting for 21%. The second-highest count is from students born in 1999.

Additionally, the effective dates of diploma and higher diploma attainment were considered, recognizing their significance as milestones toward the final degree. The number of semesters taken to complete the diploma, higher diploma, and degree was determined based on the effective dates.

Utilising the Rank tool in the Orange software and considering the significance, the features selected for the model include the number of semesters to achieve the effective date of the diploma and higher diploma, student district, birth year, z-score, OL maths grade, gender, and employability status.

The data type of our target is numeric, with values such as 7, 8, 9, and so on. Until the 19S1 semester, the number of semesters students have taken to complete the degree is documented. So that can be utilised to align with other features for constructing a predictive model. An additional feature in our dataset is the graduate status, which has binary values of 0 or 1. A status of 1 indicates that the student has completed the degree, while a status of 0 indicates he or she has not. This feature is introduced into our model to determine whether the student is a graduate or not. By analysing these selected features along with the target variable, the aim is to forecast the number of semesters new students will require to complete the degree.

3.3 File Importing and Preprocessing

For data importation, the Orange data mining tool is utilised, by dragging the file onto the workspace and selecting the necessary file. The data types could be modified through a drop-down menu, facilitating adjustments to metadata, numerical, or categorical data as needed. The File import tool allows the selection of features and the omission of unnecessary features as required. Since only the pertinent features were attached to the training dataset, there was no need to exclude any features. Once all the selected features were double-checked for their data type, role, and values, the next step was undertaken.

For the second step, only the students who have completed the degree from the dataset wanted to be selected. The 'Select Rows' tool facilitates to choose relevant records, offering built-in conditions such as 'is equal to', 'is greater than', 'is between', and so on. Using the 'Select Rows' tool, the condition 'graduate status is equal to 1' was applied to filter only the students who have completed their degree from the initial dataset. Then only the graduated students are selected to proceed to the next level.

If the target was not selected during the initial import of the file, the 'Select Columns' tool can be employed to choose the target, features, and metadata. The semester of achieving the diploma, semester of achieving the higher diploma, gender, OL maths grade, z-score, birth year, district, and employability status are identified as important features for the dataset to develop a more fitting model. The students' index number and name are also included in the dataset as metadata and will not be considered as features during the model training. The target in this context is the semester in which the degree was achieved, and it is correlated with these features to construct a more accurate model. The 'graduate status' is disregarded here since only the students who have completed their degree are selected at this stage.

When the data is selected for the model, the involvement of data sampling is considered. As the data sampling type, the tool provides access to various types such as fixed proportion of data, fixed sample size, Cross validation, Bootstrap. For our data set the fixed proportion of data is considered with 75% of the population. Since the fixed proportion of data is considered even how much the data set is, the model takes a fixed proportion of data for the inclusion in the sample. Also replicable deterministic sampling and stratified sample is available. Since a criterion is not conclusive, the stratified sample is not used for the dataset. But for the data

set the replicable sampling is selected. In every iteration a similar result should be the outcome. From this the validity and consistency of our findings can be ensured.

In 2018, students submitted their details including birthdate, Advanced Level (AL) results, Z score, etc., by completing a form and submitting it to the Centre for Open and Distance Learning. However, when students provide this information by hand, strict adherence to required fields can only be ensured if the receptionist checks them. Additionally, there may be misarrangements and displacements in some fields when information is handwritten, raising questions about accuracy. The data is not validated before being handed over to the department. After several years, the papers are then started to be inputted into the databases. Consequently, missing values emerge in such cases. It is preferable to complete the dataset before commencing modelling, as this resolves incompleteness. This necessitates performing check-ups and processes to handle missing values, incorrect values, and misarrangements. One approach involves removing rows containing missing values, while another involves replacing missing data with random values. Alternatively, missing values can be substituted with either the average or most frequent values. In our dataset, the third option, a straightforward approach wherein missing values are replaced with either the dataset's average or the most frequent values, has been chosen. Imputing missing values can mitigate biases in certain features stemming from incompleteness.

Normalisation also plays an important part in data preparation. In our dataset, according to Sri Lanka's curriculum, the Advanced Level qualification is evaluated by considering the Z-score. The AL Z-score is determined by dividing the difference between the actual mark and its mean by the standard deviation. (AL Z score calculation). In this dataset, the Z-score values range from 0.000 to 2.7564. Additionally, if the birthday is considered, it consists of an integer with 4 digits. Combining these two features onto a single pane poses a challenge. To align the numerical features within the same scale and reduce the distribution across the Cartesian plane, normalisation is utilised. In this case, Z-score standardisation is employed, with a mean (μ) of 0 and a variance (σ^2) of 1, known as standard normalisation. The following equation is used to calculate the normalised values:

$$Z = \frac{X - \mu}{\sigma}$$

Here, X represents the individual value of the specific feature.

μ denotes the mean or average value, calculated by dividing the sum of the individual values by the total number of values.

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

σ^2 represents the variance, calculated using the following formula:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$$

Here, X_i represents the individual value of the specific feature, μ is the mean, and n is the total number of values. Given that the tool offers built-in functions for calculating standard deviation, there is no need for manual computation of each equation at this stage.

Index No.	Name	Gender	OL Maths	Z Score	Birth Year	District	Employed	Diploma achieved semester	H.D. achieved semester	Degree achieved semester
E1841008	Amarasena GCS	Male	A	0.0602	1995	Hambanthota	Not Employed	2	4	7
E1841016	Ariyarathne MUP	Male	A	0.2915	1997	Kalutara	Not Employed	2	4	8
E1841045	Fernando WMTD	FeMale	A	0.2354	1997	Kalutara	Not Employed	2	4	7
E1841048	Galabada GKDDC	FeMale	A	0.0087	1999	Colombo	Not Employed	2	4	7
E1841049	Gangadeera MBTD	Male	B		1995	Galle	Not Employed	5	8	8
E1841057	Herath HMTD	FeMale	C	0.2233	1991	Kandy	Employed	2	6	8
E1841058	Inshaf IM	Male	C	0.1045	1998	Ampara	Not Employed	2	4	7
E1841061	Jayathissa KYGVP	Male	C	0.0240	1993	Nuwaraeliya	Employed	2	4	7
E1841068	Kalhara HHN	Male	A	0.0409	1997	Matara	Not Employed	5	6	8
E1841071	Karunadhara PALS	Male	C	0.0000	1999	Gampaha	Not Employed	3	4	7
E1841083	Madhusanka KWD	Male	C	0.5491	1995	Monaragala	Not Employed	4	4	7
E1841084	Madushika AK	FeMale	B	0.6179	1995	Ratnapura	Not Employed	2	5	6
E1841086	Mendis DLK	FeMale	B	0.9739	1993	Galle	Employed	2	4	7
E1841089	Mohamed Ilham S	Male	C	0.3155	1998	Nuwaraeliya	Not Employed	5	7	9
E1841093	Naveedth Ahamed AB	Male	C	0.5618	1998	Ampara	Not Employed	2	6	8
E1841105	Piratheeba G	FeMale	C	0.0179	1996	Jaffna	Not Employed	2	4	9
E1841107	Pradeetha YCM	Male	B	0.3725	1997	Kalutara	Not Employed	2	5	6

Figure 7: Before Normalisation

Index No.	Name	Gender	OL Maths	Z Score	Birth Year	District	Employed	Diploma achieved semester	H.D. achieved semester	Degree achieved semester
E1841008	Amarasena GCS	Male	A	-0.7100	-0.3916	Hambanthota	Not Employed	-0.4856	-0.5174	7
E1841016	Ariyarathne MUP	Male	A	-0.2586	0.1469	Kalutara	Not Employed	-0.4856	-0.5174	8
E1841048	Galabada GKDDC	FeMale	A	-0.8106	0.6854	Colombo	Not Employed	-0.4856	-0.5174	7
E1841049	Gangadeera MBTD	Male	B	0.0000	-0.3916	Galle	Not Employed	2.8880	3.2766	8
E1841057	Herath HMTD	FeMale	C	-0.3917	-1.4687	Kandy	Employed	-0.4856	1.3796	8
E1841058	Inshaf IM	Male	C	-0.6236	0.4161	Ampara	Not Employed	-0.4856	-0.5174	7
E1841068	Kalhara HHN	Male	A	-0.7477	0.1469	Matara	Not Employed	2.8880	1.3796	8
E1841071	Karunadhara PALS	Male	C	-0.8275	0.6854	Gampaha	Not Employed	0.6389	-0.5174	7
E1841084	Madushika AK	FeMale	B	0.3784	-0.3916	Ratnapura	Not Employed	-0.4856	0.4311	6
E1841086	Mendis DLK	FeMale	B	1.0733	-0.9302	Galle	Employed	-0.4856	-0.5174	7
E1841089	Mohamed Ilham S	Male	C	-0.2118	0.4161	Nuwaraeliya	Not Employed	2.8880	2.3281	9
E1841105	Piratheeba G	FeMale	C	-0.7926	-0.1224	Jaffna	Not Employed	-0.4856	-0.5174	9
E1841107	Pradeetha YCM	Male	B	-0.1005	0.1469	Kalutara	Not Employed	-0.4856	0.4311	6

Figure 8: After Normalisation

Figure 7 displays the data prior to normalisation, while Figure 8 illustrates the data post-normalization. If metadata such as index number and name is considered, features like Z-score, birth year, achieved diploma semester, and achieved higher diploma semester are shown as normalised. This ensures that all variables are aligned to the same scale, ranging from 0 to 1.

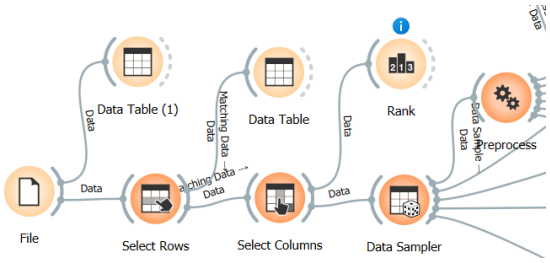

Figure 9: Data selection & Data preprocessing

Figure 9 depicts the diagram used for preprocessing the data prior to modelling.

3.4 Modelling

The subsequent step involved constructing the machine learning model, which was accomplished utilising the Orange tool. Within the tool, various built-in algorithms were available. The selection of the most appropriate model for our dataset was pivotal. Given the objective of predicting numerical values, the focus was on constructing a regression analysis model. The available built-in models, including Random Forest, Neural Network, and Linear Regression, were evaluated. The objective was to fine-tune their parameters and develop a model that most accurately defines the dataset.

3.4.1 Neural network

In a neural network, as the number of layers increases, the model's accuracy typically improves, which mitigates underfitting. However, if the number of layers exceeds the necessary amount, the model may attempt to capture all characteristics unnecessarily, leading to lower accuracy and overfitting. Therefore, it is crucial to select an appropriate number of layers that best fit the dataset. In this study, a neural network algorithm with 100 hidden layers is utilised to construct the model.

The Rectified Linear Unit (ReLU) is employed as the activation function for the dataset, defined mathematically as follows:

ReLU (x) = max(0,x)

This implies that if the input value is positive, the ReLU function outputs that value, and if the input value is not positive, it results in zero. The advantage of the ReLU function is that it does not encounter the vanishing gradient problem experienced by activation functions such as tanh and sigmoid when the gradient becomes too small.

A solver is employed as the optimization algorithm to adjust parameters such as weights and biases during the training of the dataset. This optimization function is utilised to minimise the disparity between the actual value and the predicted value. Adaptive Moment Estimation, known as Adam, is utilised as the optimization technique in this study. While balancing adaptive learning rates for every parameter, the algorithm calculates the gradient and squared gradient while considering the historical behaviour of the gradient.

The number of iterations is set to 200 so that the model undergoes forward and backward passes 200 times. Additionally, replicable training is employed, ensuring that these parameters are reproducible.

3.4.2 Linear regression

In our study, the "fit intercept" option is enabled, allowing both the intercept and gradient to be estimated by the linear regression model.

Regularisation which involves adding a penalty to the loss function to prevent overfitting is also an option here. This penalty encourages the model to be simpler, thus avoiding unnecessary complexity. However, in our case, regularisation is not employed because the features in our domain are deemed perfectly relevant. Furthermore, to effectively demonstrate the relationship between students' demographic data and their academic parameters, a simple model is preferred.

3.4.3 Random forest algorithm

In the Random Forest algorithm, the number of trees is specified as 10. Consequently, the model comprises 10 decision trees, each trained independently on the training dataset. The parameter "min_samples_split" is set to 5 in this context. "Min_samples_split" denotes the minimum number of samples required to split a node during the construction of a decision tree. As a result, if the number of samples in a node is equal to or less than 5, the decision tree will refrain from further partitioning the node.

3.5 Test and Score

Efficiency testing of various models was conducted by directly connecting the test tool to our model. Evaluation metrics including Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R2) were utilised. The training set size was configured to 75%, and the model was assessed on the test dataset.

Model	MSE	RMSE	MAE	R2
Random Forest	0.129	0.360	0.290	0.837
Neural Network	0.380	0.617	0.453	0.520
Linear Regression	0.271	0.521	0.398	0.658

Figure 10: Test scores by model

Approximately the same errors appeared after manually calculating the Mean Squared Error (MSE) for these models, confirming the values obtained from the Orange data mining tool.

Table 2: Errors manually calculated by model

	Random Forest	Neural Network	Linear Regression
$\sum Error^2$	7.1994	17.8790	12.7417
n	47	47	47
MSE	0.1531	0.3804	0.2711

All of these models aim to predict the number of semesters required for students to complete the degree program. Among them, the Random Forest algorithm was identified as the optimal model, with 10 decision trees and a min_samples_split of 5.

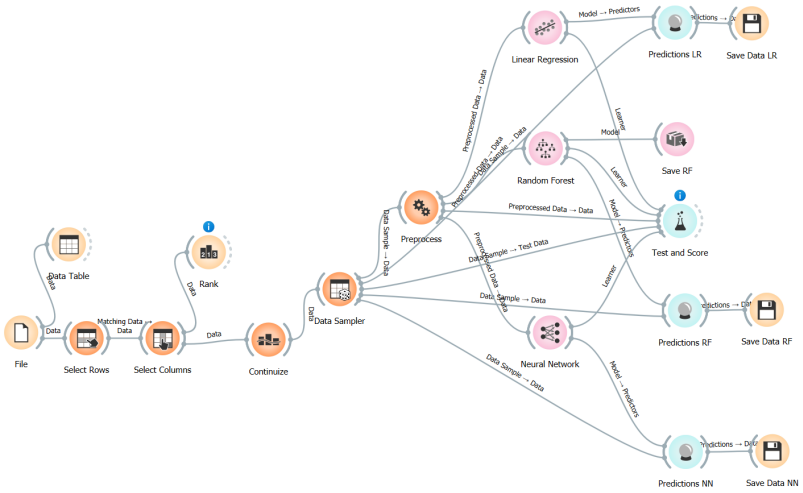

Figure 11: The diagram designed for the model building and testing

Figure 11 depicts the diagram used to design the entire modelling process.

4. Prediction

As previously mentioned, the training dataset consisted of students who graduated from the 18S1, 18S2, and 19S1 batches. Additionally, two more batches, 19S2 and 20S1, who had completed up to the higher diploma level, were included for predictions. Consequently, predictions were made regarding the number of semesters they would take to complete the degree. In August 2024, another convocation ceremony was held, resulting in the graduation of additional students from the 19S2 batch.

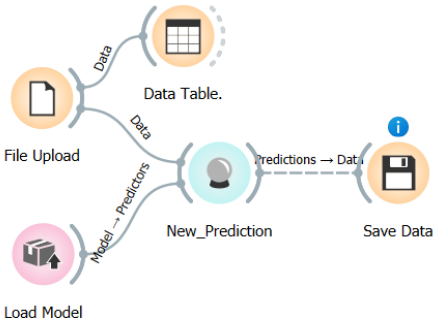

Figure 12: Diagram for prediction purpose

Figure 12 depicts the diagram utilised for prediction purposes. The process involves uploading the new dataset from the file option, followed by loading the model that best represents our dataset from the load model option. Subsequently, prediction is conducted, and the resulting output is saved in Excel file format using the save data option.

After predicting the completion duration for the 19S2 batch, a comparison was made between the predicted values and the actual values. The Mean Squared Error (MSE) was calculated as follows:

$$MSE = \frac{\sum Error^2}{n}$$

MSE = 4.15/11

MSE = 0.37

5. Conclusion

In this research, demographic features such as District, Birth Year, Gender, OL Maths, Z Score, Employability, and academic performance factors including Diploma completed semester, Higher Diploma completed semester were selected. The Random Forest Algorithm was employed for modelling, utilising 10 trees and a minimum sample split of 5. Initially, a fixed proportion sampling technique was applied, encompassing 75% of the population for sample selection. The model was trained using this data and subsequently tested on test data samples from the first three batches introduced after 2018: 18S1, 18S2, and 19S1.

The Random Forest model's performance was assessed through manual calculation of Mean Squared Error (MSE), yielding a value of 0.153, which closely approximated the tool-generated MSE of 0.129. Upon model training, it was saved as a pkls file for later use. In the prediction phase, the model was loaded to forecast the number of semesters required for students to complete their degree. Subsequently, this information was utilised to determine the overall duration for degree completion.

The model's predictive capabilities were further evaluated using data from the 19S2 batch, where predictions were made for students, some of whom subsequently graduated after convocation. By comparing the actual and predicted number of semesters taken to complete the degree, a new MSE value of 0.37 was obtained, indicating a somewhat higher level of error. It was suggested that with a sufficiently large dataset, this error could be mitigated.

The significance of this model lies in its ability to forecast the duration required for degree completion, thereby enabling estimation of the future workforce entering the industry at specific times. Additionally, by considering the backward process, it facilitates determination of the number of students needed to initiate degree programs to meet desired workforce levels. Furthermore, it allows for consideration of graduate benefits and infrastructural requirements. The versatility of this model extends to its potential application across other degree programs, provided relevant datasets are available for submission.

References

"AL Z score calculation". https://studentlanka.com/2014/07/26/how-z-score-is-calculated-for-university-selection-in-sri-lanka/. Accessed 22 May 2024.

Breiman, Leo. 2001. "Random Forests." Machine Learning 45 (1): 5-32. https://doi.org/10.1023/A:1010933404324. Accessed 22 May 2024.

Dr.Edgard F. Codd. (1970), the concept of "Database normalisation".

"Evolving technology in health sector". Digital Health Market Size, Share & Trends Analysis Report by Technology (Telehealth, mHealth, Health Analytics), by Component (Software, Services), by Region, and Segment Forecasts, (2020 – 2027). Grand View Research.

Han, Jiawei, Micheline Kamber, and Jian Pei. (2011). Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann

Haykin, Simon. (1999). Neural Networks: A Comprehensive Foundation. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Montgomery, Douglas C. (2013). Design and Analysis of Experiments. 8th ed. John Wiley & Sons.

"Moodle.org". https://moodle.org/. Accessed 22 May 2024.

"National IT-BPM Workforce Survey - 2019". (2019), https://www.icta.lk/icta-assets/uploads/2023/02/NATIONAL-IT-BPM-WORKFORCE-SURVEY-2019-1-1.pdf. Accessed 22 May 2024.

Neter, John, William Wasserman, and Michael H. Kutner. (1989). Applied Linear Regression Models. 2nd ed. Homewood, IL: Irwin.

"Orange Data Mining". https://orangedatamining.com/. Accessed 22 May 2024.

"Predicting graduation rates at Doctoral Universities". Goenner, C. F., & Snaith, S. M. (2004). Predicting Graduation Rates: An Analysis of Student and Institutional Factors at Doctoral Universities. Journal of College Student Retention: Research, Theory & Practice, 5(4), 409-420. https://doi.org/10.2190/LKJX-CL3H-1AJ5-WVPE.

"Predicting student graduation - 2024", Lidya R Pelima, Yuda Sukmana, and Y. Rosmanasyah. (2024). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. (2007). Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge: Cambridge University Press.

Sir Francis Galton. (1996), the statistical techniques of correlation and linear regression (e.g., see Millar).

"Technology in Academia: Trends Shaping the Future of Higher Education" by David W. Chapman Jr., which discusses how advancements in technology are influencing teaching, learning, and research in academia.

Warren McCulloch, and Walter Pitts. (1943). Developed the first models of Neural Networks.

Yule, (1897), the term "standard error" was first used in relation to the standard error of a residual.