The Influence of Clarity on Career Choice and Academic Success in First Year Biology

Faisal Jr. Saleem
School of Computer Science
and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
2095208@students.wits.ac.za

Shalini Dukhan
School of Animal, Plant
and Environmental Sciences
University of the Witwatersrand
Johannesburg, South Africa
shalini.dukhan@wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—In this study, multiple machine learning techniques are used to understand the link between students' experience, expectations and involvement in university, and their final grades. This paper heavily focuses on six machine learning techniques and feature selection in order to determine which factors play the most influential role in determining students' academic success in first year Biology.**

***Index Terms*—Machine Learning, Models, Academics, Student Success, South Africa**

I. Introduction

Simulating data in Python entails a systematic algorithm estimating a complex phenomenon or system. As extracted from various sources, raw data always give misleading answers; hence, data screening or cleaning is vital to generate the desired outcome. This research project is concerned with determining key factors contributing to a student's academic success. The project was to assess the influence of career choice and academic success in first-year biology. After data screening, machine learning (ML) algorithms models were executed. First, the ML models on the dataset were used without feature selection. The models applied were Decision Trees, Linear Regression, Random Forest, K-Nearest Neighbors, Multi-Layer Perceptron Neural Network, and Ridge Regression, to predict the target variable, the average of all marks the student acquired throughout the year. Obtaining the results entailed executing these six models by selecting features using two methods: Backward Elimination and Regularized Trees. In the end, all these results are compared. Charts and tables are utilised to present results as it shows and provides a greater understanding of the process of applying these different techniques. Finally, it is concluded that the K-Nearest Neighbors algorithm predicts academic success the best with 15 selected features.

II. Data Screening Cleaning

There were multiple stages in the screening process. They will be explored in this section. Data was initially extracted in excel formats and imported into the Python integrated development environment (IDE) and code editors. The column's names were long with spaces; hence, the first cleanup activities were to remove the spaces from the columns and change the column names. Further screening activities continued to give better results. For example, extent_adapted_env and teaching_learning_env were 100% missing values (Table 1). The data also gave misleading information, such as the average marks max is 76 and the minimum average of 9. Therefore, it was obvious that the data needed some scaling since the difference between values was noticeable.

TABLE I
Numerical features described

	count	mean	min	max
age	115	18.7913	18	22
extent_in_career	118	1.6102	1	15
extent_adapted_uni_env	0	NaN	NaN	NaN
teaching_learning_env	0	NaN	NaN	NaN
change_current_year	103	45.6371	1	106
dont_enjoy_museum_centers	117	3.1795	1	4
visiting_museum_science_career	117	2.0171	1	4
visting_museum_science_topic	116	1.8534	1	4
prefer_science_class	113	2.708	1	4
topics_read_real_world	117	1.5299	1	4
topics_boring	115	3.1130	1	4
enjoy_lab_practical_sess	117	2.1197	1	4
avg_exam_marks	118	52.2948	9.45	76.76

Missing values per column were also evaluated, and the graph was generated to visualise the values [9]. For the ML step, there is always a need to compensate for the small sample size through a simulated dataset with bootstrapping generated using a Python package SimPy [15]. A normal probability distribution (dnorm) is applicable to make each vector point's means and standard deviation consistent [15]. The dataset was simulated to demonstrate their unbiased depiction of the parent data through a validation exercise. Figure 1 presents a graphical representation of missing values by column. The graph shows that extent_adapted_uni_env and teaching_learning_env features are 100% missing values. The subsequent action was to remove the missing features to have cleaner data. In addition, a unique feature, Q#, was dropped or deleted. Feature values like 1.23.45 are replaced with null values to make data consistent. Replacing the values with null is more helpful than guessing what the original values might have been because that will mean manipulating parent data without knowing the

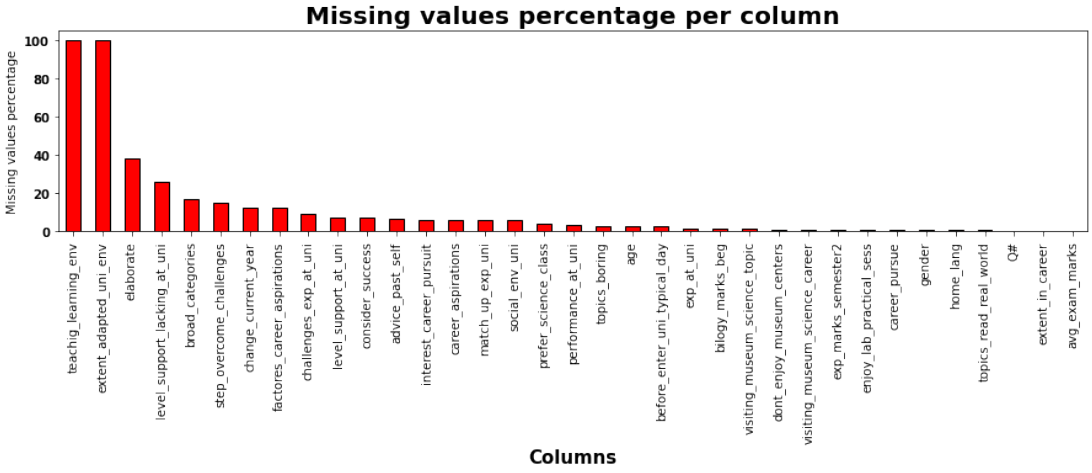

Fig. 1. Missing values percentage per feature

author's reasons for such values. The representation of the parent data should be unbiased for consistent results. Figure 2 presents samples of histograms generated, representing some feature values. Some inconsistencies in the output depict that the dataset has some outliers, which distort the outcome of the means and standard deviations.

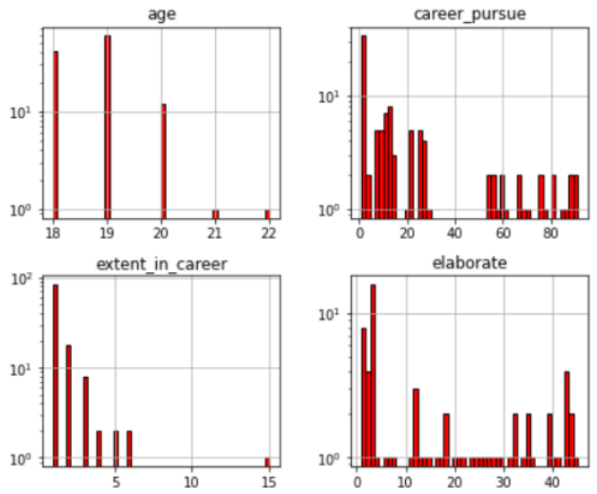

Fig. 2. Histogram of sample features

After extensive cleanup of the raw data, Table 2 presents a summary of values and information on the dataset. Firstly, the dataset was imported, assessed, and screened. Prior to imputing missing values, it was important to identify them. Figure 1 and Table 1 exhibit the missing values from the numeric features per column. Figure 2 shows distorted histograms because of various unique and missing values. There are columns with 100% missing values, which were dropped, and those with unique values. Table 2 shows that the ave_exam_marks column had 117 unique values. However, zero missing, elaborate had 40 unique values and 38% missing values, challenges_exp_at_uni had 58 unique values and 16% missing values, and advice_past_self had 71 unique values and 16.94%, to mention a few. The content of the table is self-explanatory. However, the implication is that the cleanup activity was successful, and the data is ready for model algorithm exercises. The analysis considers an in-depth investigation of significant features with missing and unique values.

This first process aimed to explore the significant properties, patterns, features, and statistics. Building on the process entailed data type conversion, imputation of missing values, column or label encoding, log transformation, and other activities. This outcome of the data screening set a fundamental basis for model building. Analytics on the success of standard features contributing to students' academic success was helpful information for administrators and policymakers to arrive at effective decisions about exam assessment. The report examined the insightful data trends hidden from students' exam academic evaluation practices, particularly for big data, such as national or international exams conducted simultaneously during the same period. The national tallying of school performance is susceptible to this kind of simulation. Drivers of academic success in various exams or movies are analysed. As such, the outcome of the report is based on reputable exploratory data analyses identifying key features and eliminating distorting values to predict outcomes and provide consistent mean and standard deviation. The original data feature depicts the average exam marks and students' academic performance on a big-data scale. The simulation is conducted to generate more values for a small sample size — the data analysis entailed classification based on missing values percentages and frequency of the unique values. Data analysis experts know that missing values analysis and screening is the core step in data processing because they could introduce inconsistencies and biased outcomes, reducing model efficacy. The percentage of missing values per variable should be used to depict possible errors and results categorised best. Classification could entail dividing the data into no missing value groups, less than five percent missing values,

TABLE II
OVERVIEW OF DATASET AFTER CLEANUP

Feature	UniqueValues	% MissingValues
elaborate	40	38.9831
step_overcome_challenges	56	28.8136
level_support_lacking_at_uni	35	28.8136
broad_categories	33	27.1186
interest_career_pursuit	58	25.4237
factor_career_aspirations	23	22.0339
typical_day_before_uni	52	18.6441
match_up_exp_uni	27	18.6441
level_support_at_uni	36	17.7966
advice_past_self	71	16.9492
consider_success	33	16.9492
challenges_exp_at_uni	58	16.1017
change_current_year	51	12.7119
career_pursue	57	12.7119
career_aspirations	17	11.8644
social_env_uni	38	8.4746
exp_at_uni	43	5.9322
prefer_science_class	4	4.2373
performance_at_uni	24	4.2373
topics_boring	4	2.5423
age	5	2.5423
biology_marks_beg	30	1.6949
visting_museum_science_topic	4	1.6949
home_lang	19	0.8475
exp_marks_semester2	42	0.8475
dont_enjoy_museum_centers	4	0.8475
visiting_museum_science_career	4	0.8475
topics_read_real_world	4	0.8475
gender	3	0.8475
enjoy_lab_practical_sess	4	0.8475
extent_in_career	7	0
avg_exam_marks	117	0

and more than five percent missing values. The categories should be further analysed in terms of the median and mode imputation and sequel imputes by 0. However, variables with more than five percent missing values should be filled with null (NA) so that they are not used for further computation because of possible distortion of the average scores.

III. MODELLING

After cleaning the data, modelling started. The feature selection using a backward approach was performed on the dataset, and then the same ML models were performed on the features selected to find the root mean squared errors (RMSE) and mean absolute errors (MAE). Feature selection is again simulated using regularized trees, and then the 6 ML models are executed on the features selected to find the corresponding model errors. The results for full features, backward feature selection, and regularized trees feature selection are compared. The results include different numbers of features comparison. Graphs have been provided that show errors vs. the number of features corresponding to methods used (Backward feature selection and regularized trees) and full features.

Backward feature selection provides high accuracy by only using 15 features selected for the K-Nearest Neighbour model. The remaining results are also shown for the other models and feature techniques. ML models have been performed on the dataset without feature selection and with feature selection. The six models used are Decision Trees, Linear Regression, K-Nearest Neighbor, Random Forest, Multi-Layer Perceptron Neural Network, and Ridge regression.

TABLE III
RMSE AND MAE VALUES FOR EACH MODEL

	RMSE	MAE
DecisionTreeRegressor	14.9478	12.3613
LinearRegression	11.7857	10.238
RandomForestRegressor	10.8878	8.4583
KNeighborsRegressor	9.9927	8.4667
MLPNeuralNetwork	39.3795	37.9378
Ridge	11.7193	10.1787

ML modelling started with encoding the categorical features. Feature scaling marks the end of data processing in ML. Encoding and scaling marks are methods used to standardise the independent variables of a dataset within a specific range. In other words, feature scaling limits the range of variables so that you can compare them on common grounds. For testing, it was necessary to split the data into 80% validation and 20% testing set. Table 3 presents the outcome of RMSE and MAE values per model.

Data transformation was conducted through a regression model prediction of the student's academic performance, and appropriate values displayed. The RMSE and MAE compared each model, evaluating the final model's performance and accuracy [2]. The analysis entails finding the correlation between a dependent feature (DF) and independent features (IF) [14]. According to Feature-engine developers [4], the data analysis relied on various feature engines (a Python library) with many transformers to select and engineer attributes for each ML model. The library applied Scikit-learn methods to evaluate parameters and transmute the dataset [4]; [13]. Pedregosa et al. [13] explain that model selection is essential for accurately determining the outcome. Scikit-learn can assess an estimator's output or particular features or parameters based on cross-validation, optionally dispensing the calculation to numerous cores. This process is accomplished by GridSearchCV object wrapping the estimator; note that the "CV" stands for the cross-validated [13]. When a parameter or code calls to fit, it chooses the specified parameter grid, optimising a score. Predict, score, or transform are then allocated to respective tuned estimators [13]. Creating and transforming data features and attributes is vital for guaranteeing the accuracy of the product and time efficacy. Model performance depends on the data processing and preprocessing effectiveness. Reddy [14] estimates that feature engineering boosts data accuracy by more than 70%. It also helps select the required IFs, and the significant IFs with more DF relations result in better model performance.

The comparison of students' academic performance was based on a dataset of exam scores evaluated based on various parameters. It was a quantitative analysis that relied on descriptive and visualisation presentation of results for straightforward interpretation [9]. According to Agrawal [1], regression algorithms or modelling determine the correlation between variables predicting discrete values. Nonetheless,

the RMSE error resulted in slight inconsistency. Using a Root Mean Squared Logarithmic Error (RMSLE) could have reduced the error margins. RMSLE is considered ideal for data analysis due to the log transformation metrics with the large scale of reducing results' error scale [1], [11]. On average, the MAE and RMSE scores were high compared to the significance or acceptable values of 0 to 1. RMSLE help generates the preferred model minus the need to call the inputs [1], [11]. In this scenario, the log of computed RMSE error would have resulted in minor inconsistency. The NumPy log function is a simple metric for encoding Big Data [1]. They are appropriate for numerical data analysis and visual result exhibition [9]. The data analysis approach helped create a model of investigated and determined results or performance of students in the exam. Interpretable or correlational models, such as linear models, predict academic performance.

The MAE assessed the mean magnitude of the errors in the performance prediction without determining the trend. It evaluates the accuracy of continuous variables. In other words, "the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation" [6]. The MAE provides a linear score depicting that weighting of variables is not discriminative of individual differences of variables. On the other hand, the RMSE is a quadratic measuring standard for the mean magnitude of the variable errors [2], [6]. RMSE is expressed in words as "the difference between forecast, and corresponding observed values are each squared and then averaged over the sample" [6]. Because the errors are squared before averaging, the RMSE returns a comparatively high weight to significant errors [11]. RMSE residual SD or prediction errors are relevant to making a judgemental conclusion; however, the significant errors depend on how well the raw data is screened. The accuracy of simulated data relies on how well the data is cleaned to remove deviating values. RMSE measures the broad residual values from the mean or the data points. It is applicable for determining how close the values concentrate toward finding the best fit through regression analysis and forecasting [11]. When standardised forecasts and observations are applied in RMSE inputs, a direct association with the correlation coefficient is found; hence, it is possible to determine students' academic performance using the RMSE algorithm in a simulated manner. Hence, it is most useful when huge errors are predominantly unwanted.

MAE depicts the difference between predicted and original values by the absolute average of the set. It is a scale-dependent accuracy because of errors in observation computation and is applicable for regression models in ML. RMSE and MAE evaluate the fitness of the algorithm. However, the former predicts the model with reverence to the projected variable model as defined by the square root of the mean squared error [15]. During the assessment, RMSE computed of 95% confidence interval (CI) while bootstrapped procedures estimated MAE values with a fixed seed (t) [15]. Each MAE float value should range from 0.0 to 1.0, with the best value being 0.0. Figure 3 presents the comparison performance of RMSE and MAE, with the former having higher values than the latter. However, each model score is higher than the floated values of either RMSE or MAE but is equally crucial for comparing the six models against students' academic performance. Figure 4 presents the scores of MAE per each model. From the results, K Nearest Neighbor had the lowest score (best model), and Multi-Layer Perceptron Neural Network had the highest score (worst model). According to Komorowski et al. [10], correlation techniques portray the modification of a variable as a function and allow easy comparison of results. Covariance and correlation evaluate random variables' correlation levels positively or negatively [10]. The variables depicted a positive correlation.

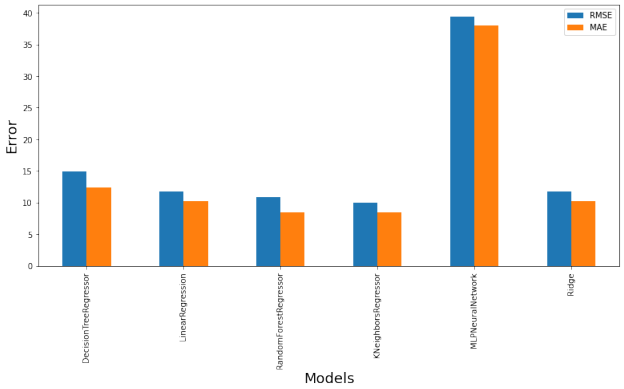

Fig. 3. Comparative Performance of RMSE and MAE

A. Decision Trees

A decision tree (DT) is essential for making decisions about data or finding based on the known probability of occurrence of different circumstances. It is intuitively applicable for probabilistic analysis because of its simplicity and ease of interpretation. Significantly it approximates complex nonlinear relationships and scales large datasets easily [10]. However, the disadvantage is rigidness and smoothness. In data analysis, it is essential to divide the space of possible predictor values into non-overlapping intervals [15]. The root nodes and internal nodes are essential features of this model. Unlike linear regression models, tree-based models need hyperparameters. Therefore, a grid search is executed to systematically go through all points of hyperparameters within the searching space and assess model performances based on varied hyperparameter integration [15]. The model was executed, and the results of RMSE and MAE were 14.94 and 12.36, respectively. The time for detection was 0.007 seconds. The scores of both RMSE and MAE are beyond the threshold; however, applicable for the scope of this study, which was finding the best fit model among the listed six. The benefits of DT revolve around the use of various forms of data without

standardization because the logistics and predictions are not sensitive to the effects of standardized data. As displayed in Table 3, this was the fifth-best model among the six because of the high RMSE and MAE scores despite taking the shortest time to execute. The execution time is important for evaluating a model because time matters and everything is gauged on the efficiency of time used. The findings of this model were compared to the output of the remaining five to decide, which the best was. Nonetheless, it is possible to say that log transformation may have generated different results.

B. Linear Regression

The OLS regression model is a controlled ML algorithm for excellent computation speed and interpretability [5]. It fits a straight line along the data points, minimizing the number of squares observed between data and forecast values. As the square root of an inconsistency, RMSE is understood as the SD of the unsolved variance and can be in the same units as the response variable [5]. Lower values of RMSE depict better fit measuring how accurately the model depicts the outcome. A low RMSE value shows that simulated and actual data are close to each other, indicating better accuracy [6]. Therefore, a lower value is better for explaining model performance and selecting a better model based on the observed mean. It is one of the directives for validating the dependability of a model [7]. The dependability of a model also extends to accuracy and its fitness to a given phenomenon. The value of RMSE, which is 0.8, is acceptable. This value is important for evaluating the efficiency of an individual model; however, when comparing different modes the log transformation is, arguably, the most popular among the different types of transformations used to transform skewed data to approximately conform to normality. If the original data follows a log-normal distribution or approximately so, then the log-transformed data follows a normal or near-normal distribution. The value generated by the system determines the comparison decision. Particularly, this was the case in this study because the scores were higher than this threshold, only the generated values were compared against each other. This model resulted in the RMSE being 11.78 and MAE being 10.24. The time taken was 0.075 seconds longer than the decision tree regression. The findings indicate that the model is linear regression was the fourth best among the selected models. It gives a statistically significant correlation between students' performance and other measures despite being higher than the threshold figure. Importantly, it is reasonable to note that the outcome could be related to data processing and log transformation aspects.

C. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a robust classification algorithm with clear intuition and mathematical detail applicable in a real-world dataset to observe how they function [8]. It is always essential to predict Direction based on the percentage returns of Lag1 and Lag2. The KNN model is built using the SciKit-learn Python library [8]. This algorithm is applicable for regression and classification to predict the results of observations and compare with k similar cases, where k is defined as an analyst. The ideal or best k reduces the RMSE prediction error [8]. The data should be standardised while performing KNN algorithm analysis. The lag aspects of the data analysis imply the processing of data in different ways, but the results are relevant and applicable to this project. For the current study, KNN was conducted or computed to predict error RMSE and MAE. This model concluded with an RMSE that was 9.99 and MAE that was 8.467 at the quickest time of 0.004 seconds. The values are the lowest among the six models; however, they are beyond the threshold or the acceptable range. The implication of the results is that the model is the best fit for predicting students' academic performance based on the data processing and the raw data provided.

D. Random Forest

A random forest combines multiple decision trees. It uses average prediction to show the results of the many decision trees. The bagging technique is familiar to this algorithm model [15]. It implies selecting subsets from various sets of uniformly arranged decision trees and adjusting the correlation between them. In this model, the maximum depth of the tree (max_depth) and the number of trees (n_estimators) are the only hyperparameters [15]. The feature significance in this model is almost the same compared to the decision tree method. Random searches were conducted to identify the optimal hyperparameter. The "Learning_rate" parameter depicts the iteration rate. A more considerable learning rate value depicts a quicker iteration speed that results in finding each leaf node faster. While a smaller learning rate tends to identify the optimal value with a slower iteration speed, consuming more algorithm space and cost. For the current case, the sample size used had no equal samples assigned to the test. Sklearn. Ensemble (Random Forest Regressor package), a Python module, was applicable in running the RF regressor. Replacement code helps in selecting [15]. NumPy and Scipy were applicable as python dependencies for running [15]. However, the scope of the study was to find the relationship between RF RMSE performance and RF MAE performance. This model shows an RMSE that was 10.88 and MAE that was 8.45 within 0.388 seconds. The results are depictive of the model performance against the five others, without considering the value threshold issue. Table 3 results show that this model was the second best model because of the second lowest scores of RMSE and MAE algorithms.

E. Multi-Layer Perceptron Neural Network

The Multi-Layer Perceptron model involves iterative, iteration, and partial derivatives of the dataset using the loss function to update various parameters. It is also possible to use regularization of the loss function to prevent overfitting in the model [12]. Many ML algorithms need quantitative data to denote categorical columns in a statistical column. MLPClassifier can help achieve the prediction; however, it must first be an imported object created using MLPClassifier

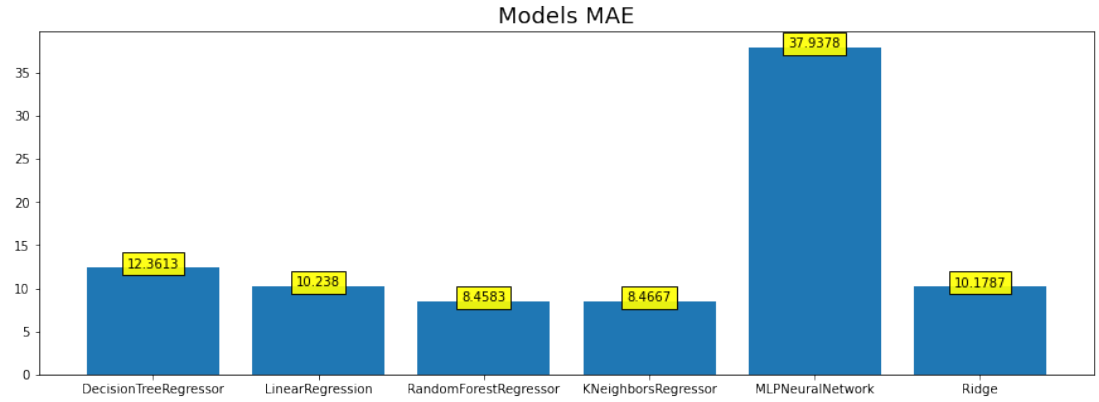

Fig. 4. MAE of Models

() code. Then, fit () is used for the model and prediction on the data set using predict () [12]. The study entailed analyzing the performance of the MLP neural network on student academic performance based on RMSE and MAE performance. This model achieved an RMSE of 39.39 (the highest among the models) and MAE of 37.94 (the highest). The classification rate was not specified or measured. The results show that this was the last best-fit model for the simulated data. However, it is necessary to say that the data processing process limitation may have contributed to the outcome. Further research is necessary for testing the outcome based on different platforms and metrics.

F. Ridge Regression

Ridge regression is one of the most popular linear regression regularization techniques. It is a viable solution for multi-collinearity in IVs [15]. A particular 'bias' is incorporated into the data to expect a better lasting forecast. In order to overcome challenges associated with this model, computing a sufficient working matrix is proposed. The parameter is approximated and calculated from the matrix of adequate statistics. The dataset loaded once, which suggestively sped up the performance [3]. RR tested students' performance in the exam, and the results presented depict whether it was an ideal model. This model shows an RMSE that was 11.71 and MAE that was 10.17 at a detection time of 0.004 seconds. Among the six models, it was the third best. Findings show that RR was the third-best model for assessing students' academic performance based on the data provided.

IV. Conclusion

Feature selection techniques are critical in predicting any target variable as precisely as possible. This research assessed students' academic performance using Python ML models. Six models were evaluated based on RMSE and MAE. Intensive data processing was first conducted for cleaning missing and unique values. RMSE and MAE low scores depict the best model; therefore, K Nearest Neighbor was the best model among the six models, and Multi-Layer Perceptron Neural Network was the most minor applicable for the analysis. On overview, none of the models provided the most optimal representation of the evaluation desired because values were above the acceptance range of MAE and RMSE. Using log transformation, particularly RMSLE could have helped find the ideal model. Predicting student performance in exams is a quantitative approach that is complex due to the volume of work; however, with simulation, it is possible to find the best fit because of available algorithms. The scores of the algorithm determine the ideal one; however, the fit may be selected based on the data set used, not necessarily the method. In this project, the six models, Decision Trees, Linear Regression, K-Nearest Neighbor, Random Forest, Multi-Layer Perceptron Neural Network, and Ridge Regression, helped to find the best fit for the student's academic results. However, the RMSE and MAE scores were higher than the standardized or acceptable ranges between 0 and 1. This outcome could be linked to data processing limitations and the restriction of the two algorithms. A possible reason for having higher values than accepted significance limits is the lack of log transformation of data. The log transformation is essential and viable for simulation because it assists to change or align skewed data to conform to normality. However, it is important that the skewed data follow a log-normal approximation and distribution. Notably, the log-transformed data conform to normal distribution. Lack of RMSLE notwithstanding, the use of RSME and MAR was a suitable analysis because the findings conformed to the required outcome of finding the best fit model among the six selected ones. It was possible to determine the best and least fit based on the RMSE and MAE scores. Machine learning technology has helped to simulate the results of students' academic performance with a high level of accuracy; hence, the results are applicable for generalisability. However, future research can assess the application of RMSLE to this data for comparison purposes.

References

[1] R. Agrawal. Know the best evaluation metrics for your regression model. 2022. https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model.

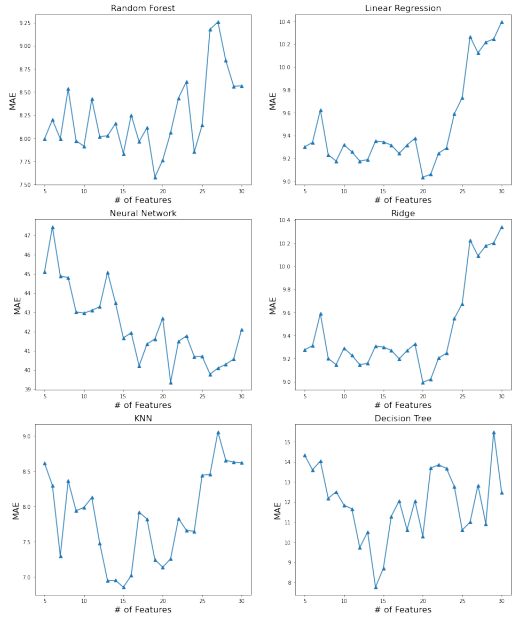

Fig. 5. Backward Elimination

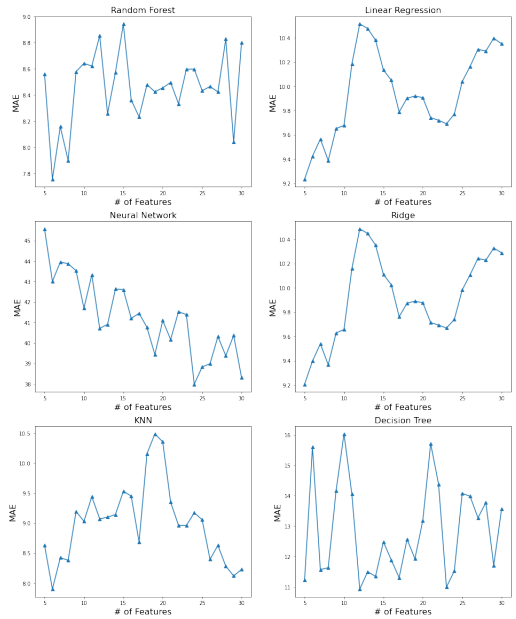

Fig. 6. Regularized Trees

[2] T. Chai and R. R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature." Geoscientific Model Development, vol. 7, pp. 1247–1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014

[3] W. Chiang, X. Liu, T. Zhang and B. Yang. "A study of exact ridge regression for big data." Conference Paper, pp. 1-11. 2018. https://doi.org/10.1109/BigData.2018.8622274.

[4] Feature-engine developers. Feature-engine: A Python library for feature engineering and selection. 2022. [Online]. Available: https://feature-engine.readthedocs.io/en/latest/ [Accessed Nov. 12, 2022].

[5] A. Gelman, J. Hill, and A. Vehtari. "Regression and other stories." Cambridge University Press, 2020. https://doi.org/10.1017/9781139161879.

[6] O. T. Hodson. "Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not." Geoscientific Model Development, pp. 1-10. 2022. https://doi.org/10.5194/gmd-2022-64.

[7] T. O. Hodson, T. M. Over and S. F. Foks. "Mean squared error, deconstructed." Journal of Advances in Modeling Earth Systems, vol. 13, no. 12, pp. 1-10. https://doi.org/10.1029/2021MS002681.

[8] A. Kassambara. Machine learning essentials: [practical guide in R]. Sthda. 2017.

[9] N. Kavi. Time series data visualization using Heatmaps in Python. 2021. [Online] Available: https://www.analyticsvidhya.com/blog/2021/02/visualization-in-time-series-using-heatmaps-in-python/ [Accessed Nov. 12, 2022].

[10] M. Komorowski, C. D. Marshall, J. D. Salciccioli, and Y. Crutain. "Exploratory data analysis." In: Secondary Analysis of Electronic Health Records, pp.185-203. 2016. https://doi.org/10.1007/978-3-319-43742-2_15.

[11] C. Lepelaars, C. (2022). Understanding the metric: RMSLE. 2022. [Online]. Available: https://www.kaggle.com/code/carlolepelaars/understanding-the-metric-rmsle/notebook [Accessed Nov. 12, 2022].

[12] A. Navlani. Multi-layer perceptron neural network using python. Machine Learning Geek. 2021. [Online] Available: https://machinelearninggeek.com/multi-layer-perceptron-neural-network-using-python/ [Accessed Nov. 12, 2022].

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. "Scikit-learn. Machine Learning in Python." Journal of Machine Learning Research, vol. 12, pp. 2825-2830. 2011. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[14] K. P. E. Reddy. "Step by step process of feature engineering for machine learning algorithms in data science." Analytics Vidhya. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/03/step-by-step-process-of-feature-engineering-for-machine-learning-algorithms-in-data-science/ [Accessed Nov. 12, 2022].

[15] S. Sazawal, S. Das, K.K. Ryckman, R. Khanam, I. Nisar, S. Deb, E.A. Jasper, S. Rahman, U. Mehmood, A. Sutta, N.H. Chowdhury, A. Barkat, H. Mittal, S. Ahmed, F. Khalid, S. M. Ali, R. Raqib, M. Ilyas, A. Nizar, … et al. "Machine learning prediction of gestational age from metabolic screening markers resistant to ambient temperature transportation: Facilitating use of this technology in low resource settings of South Asia and East Africa." Journal of Global Health, vol. 12, no. 0402, pp. 1-11. 2022. https://doi.org/10.7189/jogh.12.04021.