University of the West of England BRISTOL

Student ID: 24004615

Module Number: UFCFMJ-15-M

Module Name: Machine Learning

Word Count: 2280 Words

Referencing: Harvard Referencing

3

The aim of this assessment is for you to design your own ‘smart thing’ and describe this in a report

Evaluation of Predictive Models for Student Dropout Prediction

Introduction

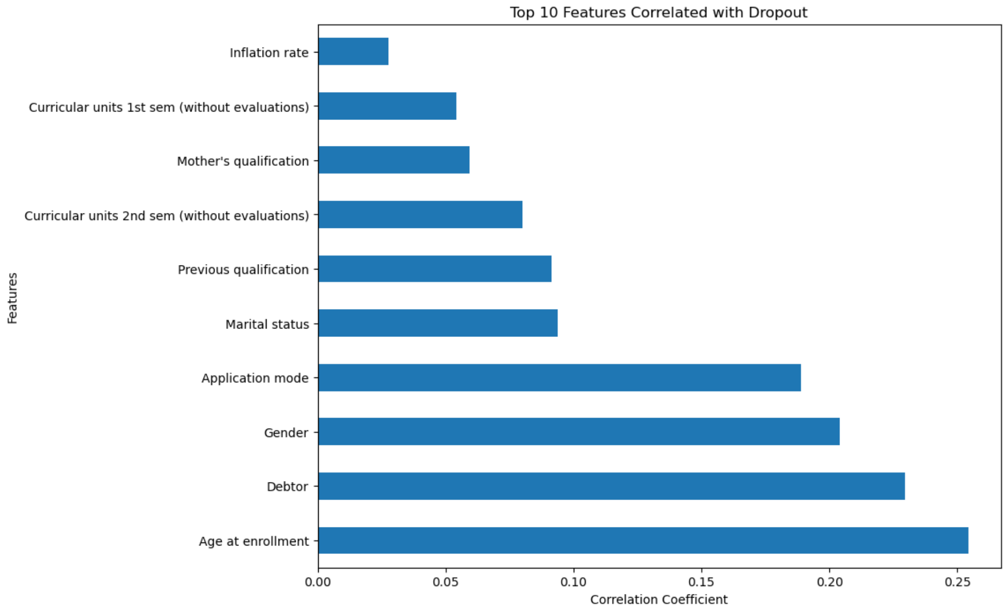
The advancements in field of machine learning (ML) have impacted various sectors significantly. For instance, in healthcare sector, these ML models are deployed to predict various diseases based on clinical symptoms. According to research study by (Duarte et al. 2018), predictive models have been deployed in clinical settings to detect patients who're at risk of heart failure. Similarly, in educational settings, predictive models have been deployed to enhance educational evaluation through predictive student assessment modelling (Pham Xuan Lam et al., 2024).

Particularly, predictive models have become essential in making informed decisions across all industries. By analyzing the existing data, these models enable organizations to manage risks, forecast future outcomes, enhance customer experience and optimize resources (Sghir, Adadi and Lahmer, 2022). As educational institutions are widely adopting ML, predictive models are playing important role in comprehending student behavior, making retention rate better and ensuring academic success (Cui et al., 2019).

This study explores the use of predictive modeling to identify risk of student dropouts at an early stage, providing chance to university to intervene proactively and implement strategies to support them. Early identification enables universities to impose targeted support for example, customized learning plan, tutoring resources, academic and mental health counselling to assist students who're at risk of dropping out.

Exploratory Data Analysis

The dataset used to train and evaluate ML predictive models is acquired from publicly available resources from UC Irvine machine learning repository. The dataset it built upon study conducted at institution of Portugal on students during academic path consisting of their enrollment dates, demographics and social economic factors to analyze the student outcomes. Data consists of 34 different features represented in form of columns and 4424 rows represents unique student. The dataset has multiple data types such as binary, categorical and continuous. The most important feature of data is 'Target' which is categorical in nature and formulated as three category classification tasks i.e., Enrolled, Dropout and Graduated at the end of normal duration of the course.
During initial findings while conducting exploratory data analysis it is observed that the number of dropouts is higher than the number of graduates as shown in figure 1.1 and 1.2.



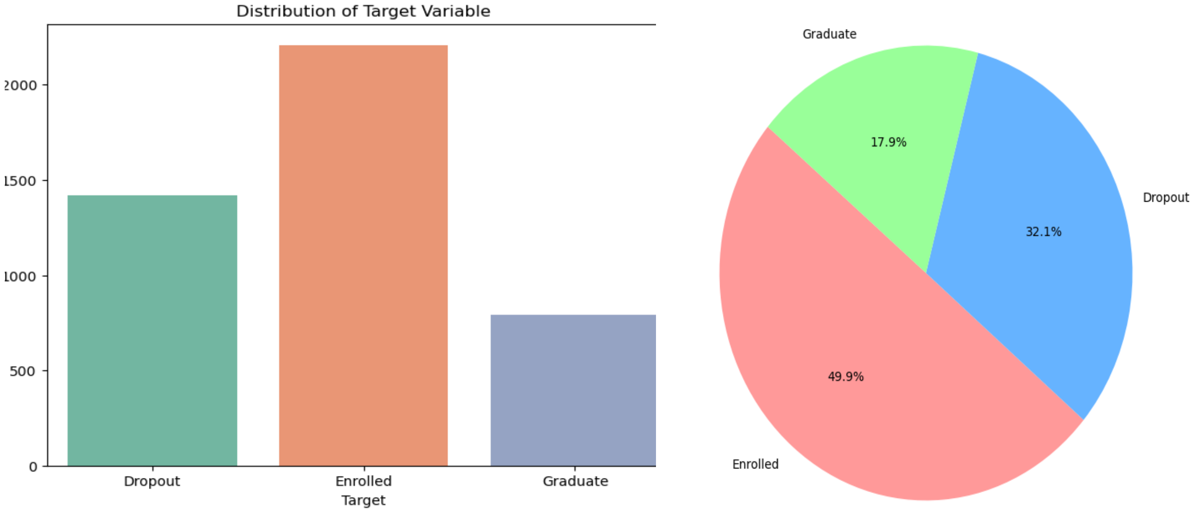


Figure 1.1 & 1.2

Following section will evaluate the features which directly influence target variable i.e., student outcomes.

Figure 2

Figure 2 illustrates top 10 features that have the strongest correlation with student outcomes, using Random Forest correlation coefficients. The strongest correlation feature are: Age at enrollment has correlation of approximately 25% with target variable. This could be due to several reasons such as returning to education after long gap, increased responsibilities and work-life balance. Debtor, with significant correlation of ~23%.any kind of financial stress can increase the risk of student dropouts. Furthermore, Gender has an evident correlation of ~20% which may be due to societal expectation contributing to risk of increased dropouts. Application mode shows some positive correlation of approximately ~19%, which can be interpreted as some application modes may link to various student demographics that can influence student dropout rate. While Marital Status and Previous Qualification have mild relation to dropout rate with correlation of 9%. For instance, some students who have responsibility of dependents may find it difficult to focus on their education which can become a reason for increased dropout risk Students with not sufficient education background may get difficulty in higher education which is demanding are likely to dropout. Furthermore, Curricular Unit 2nd Sem (without evaluation) have correlation estimated around ~8%, interpreted as students who do not have evaluation in second semester are likely to dropout due to academic struggle or disengagement. Mother's Qualification, Curricular Units 1st Sem (Without Evaluations) and Inflation Rate have correlation of roughly 6%, 5% and 3% respectively. These features show some positive correlation but do not have significant enough to create an impact on student outcomes.

As a result to initial exploratory data analysis, it can be concluded as shown in figure 2 that student outcomes are strongly associated to age, financial stress and early academic performance. These insights suggest that higher educational institutes can provide targeted support for older students, early academic assistance and financial aid for students to reduce overall dropout rates.

Evaluation Metrics

To analyze the performance of any predictive model for student dropout and academic success, several evaluation metrics are considered such as accuracy, precision, recall, F-1 score and cross validation accuracy. Theses parameters of evaluation metrics give comprehensive knowledge about how well the model is preforming and predicting outcome, specifically in context of multiclass classification problem (Oona Rainio, Jarmo Teuho and Riku Klén, 2024).

The basic principle of calculating accuracy is by calculating portion of correct prediction from total number of predictions.

$$Accuracy = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions}$$

Accuracy provides an extensive sense of predictive model's efficiency; however, it can be deceptive in case of uneven data where dataset is dominated by one class. Although accuracy is important metric, however relying merely on this parameter might be insufficient in accurate prediction of dropout cases.

Precision is the measure of true positive prediction made by the model. And is calculated as ratio of correctly predicted cases to total predicted positive cases.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Precision is important in situations where the cost of false positives is significant. In current scenario, if it is anticipated that students will dropout school (a positive case) but they do not, it may result in unnecessary interventions

Recall is the ability of model to measure all actual positive cases. It is formulated as the ratio of correctly predicted positive to all cases that are actually positive.

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

Recall is specifically crucial in cases where consequences of missing positive cases are serious. For student dropout, having high recall means the model can detect large number of students who may drop out, enabling timely interventions.

The F1-Score is harmonic mean of Precision and Recall, providing a single metric that balances trade-off between them.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

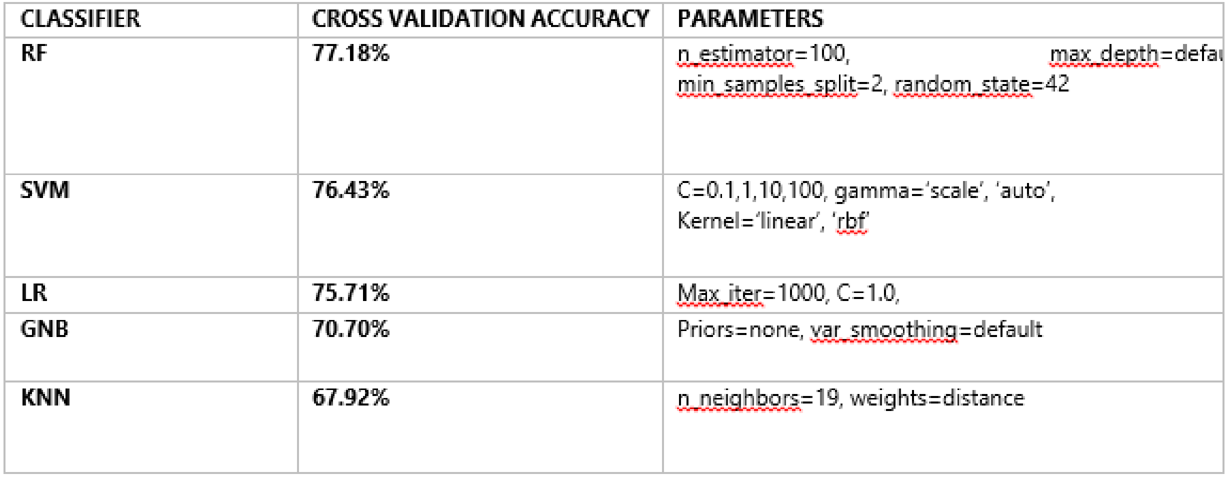
The F1-Score is especially valuable in cases of uneven class distribution, as it addresses trade-off between false positives and false negatives. In this research, F1-Score is used to make sure the model predicts dropouts accurately and in a balanced way, reducing both false alarms and missed cases.

Cross-validation was utilized to make sure that the model's adequacy remains stable and is not dependent on a specific division between training and testing datasets. During this procedure, the data was divided into five portions, with the model being trained on four of them and then tested on remaining one, iteratively (Refaeilzadeh, Tang and Liu, 2009).

Cross-validation reduces variance from train-test split randomness to provide reliable model performance estimate. This step was essential in confirming the reliability of predictive model's accuracy and other metrics are reliable and generalize towards unfamiliar data.

Predictive Models

During assessment of student dropout and academic success prediction, multiple predictive models were used to evaluate their classification accuracy. Following is a breakdown of the models utilized, listed by their accuracy ratings, along with an analysis of how well they performed.



CLASSIFIER	CROSS VALIDATION ACCURACY	PARAMETERS
RF	77.18%	n_estimator=100, max_depth=defau min_samples_split=2, random_state=42
SVM	76.43%	C=0.1,1,10,100, gamma='scale', 'auto', Kernel='linear', 'rbf'
LR	75.71%	Max_iter=1000, C=1.0,
GNB	70.70%	Priors=none, var_smoothing=default
KNN	67.92%	n_neighbors=19, weights=distance

The table above presents a clear picture about predictive models' performances and parameters selected. Random forest model presents the best performance, by depicting best accuracy rate both in testing and cross validation. Furthermore, models like logistic regression and random forest observe less difference between test accuracy and cross validation, as these models are less susceptible to overfitting. Whereas models like KNN and GNB have low accuracy indicating these classifiers are not suitable for current dataset. The following figure shows the comparison between test accuracies and cross-validation accuracies of selected ML models.

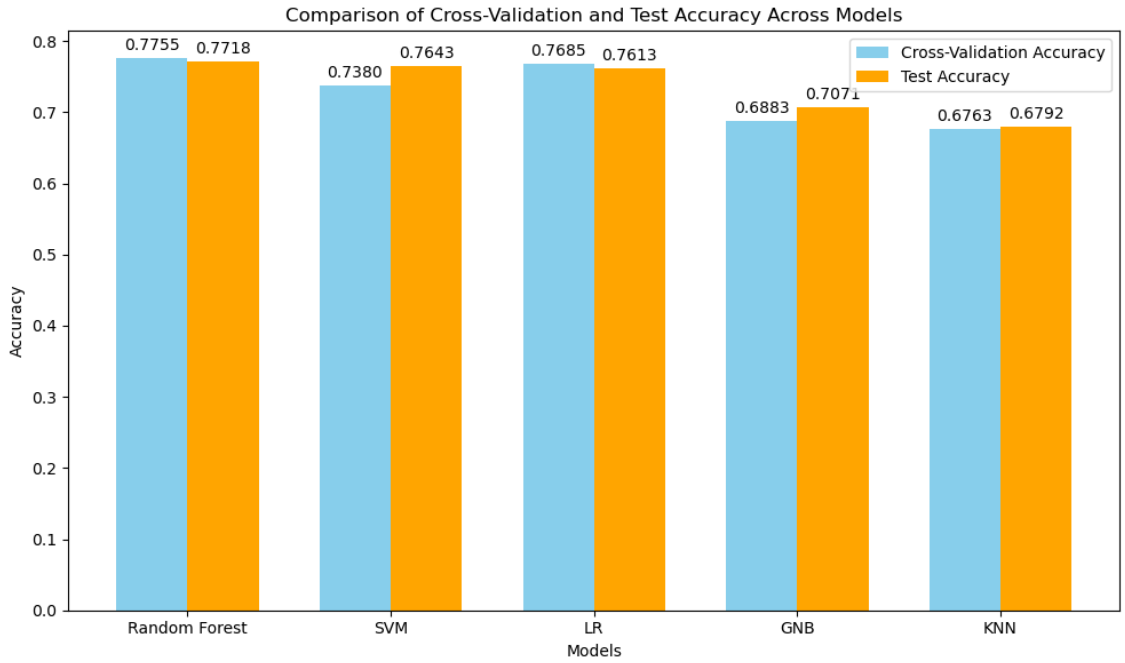


Figure 3

1. Random forest

Random forest achieved highest accuracy with 77.18% test results making it most reliable model in prediction of student outcomes. Random Forest is technique in ensemble learning where several decision trees are built during the training process to determine most commonly occurring class in classification. Random Forest reduces the chance of overfitting by averaging the outcomes of several decision trees, especially important for handling complex datasets. Random Forest offers information on most vital features for making predictions, aiding in the comprehension of factors that have greatest impact on student success or dropout rates as shown in following figure 4. Random Forest was chosen for its balance of performance, interpretability and its exceptional accuracy and capability to manage complex, high-dimensional data.

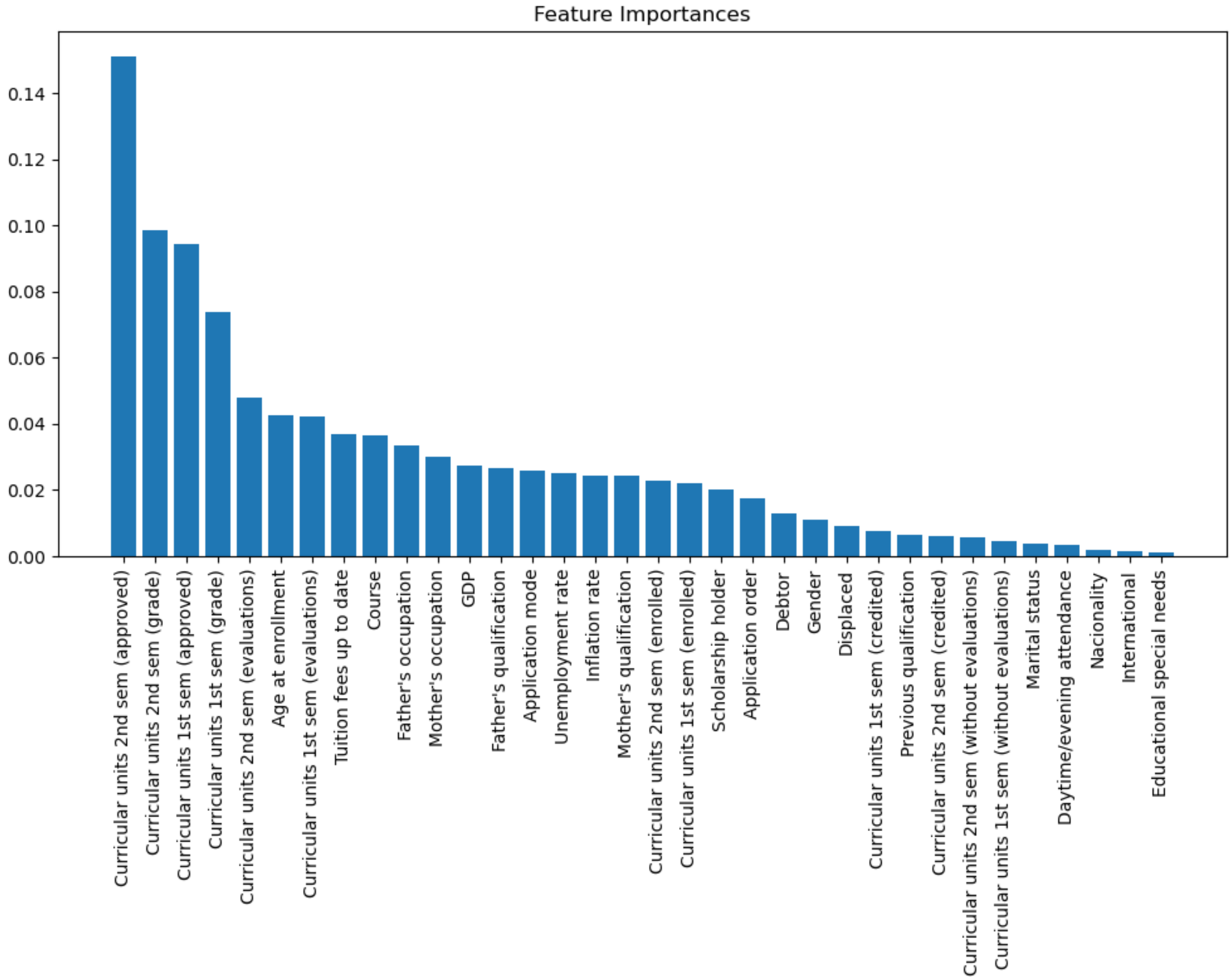


Figure 4

2. Support vector Machine (SVM):

SVM achieved high accuracy closer to random forest of 76.43%, illustrating its efficiency in classification tasks. SVM performs well in datasets with high number of features compared to the number of samples. By employing kernel tricks, SVM can efficiently divide classes in datasets with non-linear feature relationships. SVM is an effective classification technique that operates by determining the most suitable hyperplane for dividing data into distinct categories. SVM was selected for its high effectiveness and capability to manage complex, non-linear connections within the data, making it an ideal option for detailed classification tasks such as forecasting student results.

The following figure 5 is evidence of effect of hyperparameter ‘C’ on SVM at the beginning where cross validation is low, indicating model was under preforming. This is due to small value of C which allows larger margin by increasing tolerance for misclassification. Initially this factor led to underfitting. As value of C increases to 10^1 accuracy is increased prominently, Cross Validation Accuracy achieved it highest level implying that model is performing better at predicting data without over fitting. Implying that the SVM model is maintaining equilibrium between classification accuracy and margin width.

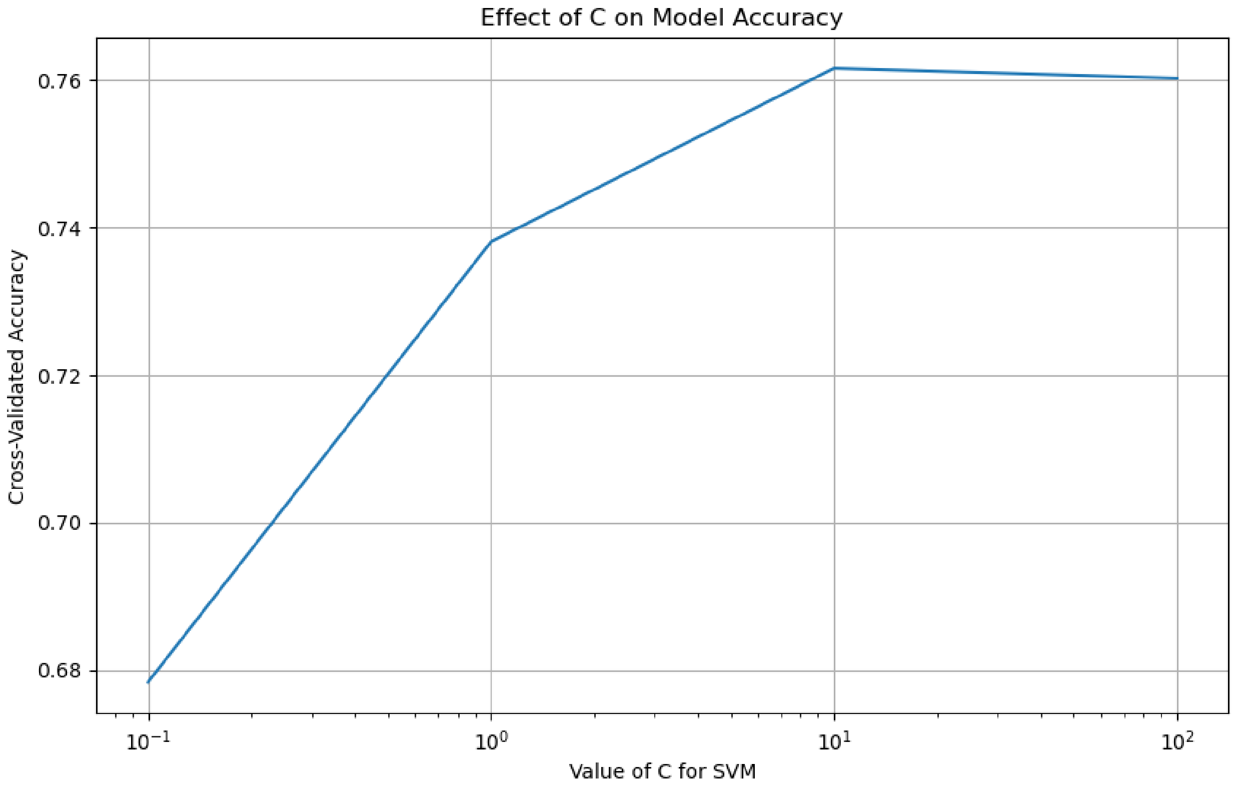


Figure 5

3. Logistic Regression

Logistic regression is simple yet powerful predictive model. It shows strength and direction of each predictor's influence on outcome and provides coefficient which can easily be interpreted. In current scenario, logistic regression model performed efficiently with an accuracy of 75.71% offering solid baseline that is easy to understand and implement. Mostly this regression model is utilized for binary classification but sometimes can be implemented to multi class problems in current study. Although logistic regression model is a simple linear model, it was selected due to its straightforward interpretation and relatively good performance, establishing it as dependable foundational model. The following figure 6 shows ROC curve for logistic regression model, representing classes 0, 1 and 2 (dropout, enrolled and graduate) in blue, red and green respectively. The area under curve shows a numerical value, providing summary of the model's performance. On the y-axis, scale goes from 0 to 1, where 1 represents optimal performance and 0.5 represents classification equivalent to random guessing. As depicted in graph, the linear regression model shows excellent ability to classify class 0 and 2, whereas class 1 shows satisfactory ability of prediction.

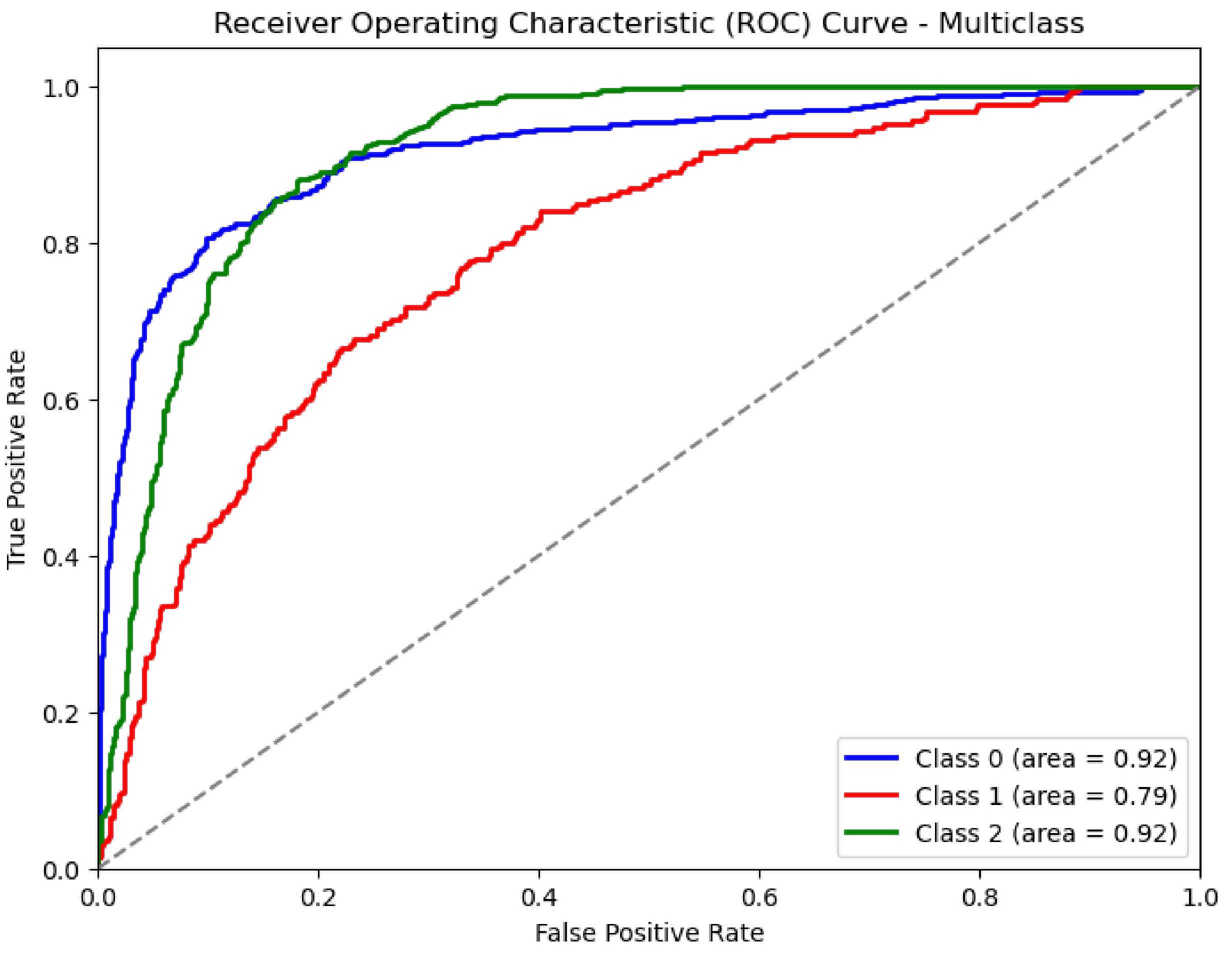


Figure 6

4. Gaussian Naive Bayes

Naive Bayes is a statistical model that follows Bayes' theorem and assumes feature independence when given class label. The Gaussian variant assumes that features follow normal distribution (Kumar Bhowmik, 2015). It is mostly suitable when the dataset is smaller, less complex or its assumptions are independent which is very unlikely in real life. Due to its low computational demand, this model was evaluated for current dataset giving an accuracy of 70.71% which is lower than the other selected ML models making it unsuitable for predicting accurate student outcomes. Furthermore, this algorithm is often criticized for its zero-frequency problem as it assigns zero probability to categorical variable whose category in test data was absent during training making it unreliable for deployment.

5. K-Nearest Neighbors (KNN)

KNN is a type of instance-based learning algorithm that is non-parametric and determines the classification of a sample by considering the majority class of its closest neighbors. KNN is easy to interpret however, it has given lowermost accuracy of 67.92% which may be attributed to curse of dimensionality, as KNN performs poorly with high feature numbers or noisy data. To improve performance of this predictive model, precise tuning of both the number of neighbors (K) and distance metrics is required. The following figure 7 shows accuracy of KNN with respect to value of K. Initially it can be rapid increase in accuracy is observed as value of k moves which might be because of model's sensitivity towards noise in data resulting in low accuracy. Subsequently, a sudden dip is observed as value of K reaches to two suggesting that certain value is not good for the data, probably because of sensitivity towards nearest neighbor or overfitting. As the value moves beyond five and reaches region between 10-15 stability in accuracy can be observed. This finding implies that at this point best balance between variance and bias is achieved giving the highest model accuracy. Beyond this point accuracy fluctuates suggesting increasing neighbors do not give better performance (Mucherino, Papajorgji and Pardalos, 2009).s

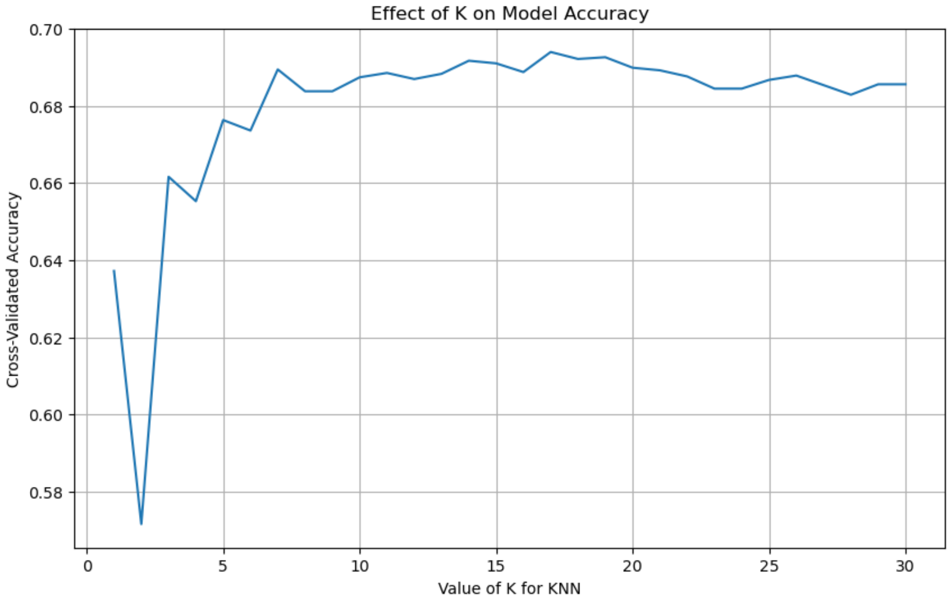


Figure 7

Conclusion

This study evaluated performance of selected ML models based on various evaluation metrics for their ability to predict student outcomes in higher education. The choice of these techniques was influenced by their combination of precision, ease of understanding, and the unique characteristics that align with challenges of predicting student dropout and academic achievement. Random Forest and SVM presented high accuracy making them a robust model in predicting student outcomes. While Logistic Regression was preferred for its interpretability. Gaussian Naive Bayes and KNN were evaluated to their simplicity and efficiency and proved to be insufficient for predicting multi-class problem. In this scenario, models that provide better accuracy such as Random Forest and SVM would be more suitable for practical use in making predictions in real-time effectively. However, for effective interpretation, staff training would be required if they’re to deployed in an educational setting.

To further improve the prediction accuracy, these selected models can be fine-tuned to achieve desired outcomes. Additionally, deep learning models and complex ensemble models can be explored and evaluated, which may require additional computational resources and hyperparameter tuning to achieve high prediction accuracies.

References

1. Duarte, K.; Monnez, J.M.; Albuisson, E. Methodology for constructing a short-term event risk score in heart failure patients. Appl. Math. 2018, 9, 954–974. [Google Scholar] [CrossRef] [Green Version]

2. Pham Xuan Lam, Mai, H., Quang Hung Nguyen, Pham, T., Hong, T. and Thi Huyen Nguyen (2024) Enhancing Educational Evaluation through Predictive Student Assessment Modeling. Computers and education. Artificial intelligence [online]. pp. 100244–100244.

3. https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

4. Sghir, N., Adadi, A. and Lahmer, M. (2022) Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). Education and Information Technologies [online]. 28.

5. Cui, Y., Chen, F., Shiri, A. and Fan, Y. (2019) Predictive analytic models of student success in higher education. Information and Learning Sciences [online]. 120 (3/4), pp. 208–227.

6. Oona Rainio, Jarmo Teuho and Riku Klén (2024) Evaluation metrics and statistical tests for machine learning. Scientific Reports [online]. 14 (1)].

7. Refaeilzadeh, P., Tang, L. and Liu, H. (2009) Cross-Validation. Encyclopedia of Database Systems [online]. pp. 532–538. Available from: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_565.

8. Kumar Bhowmik, T. (2015) Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. INTELIGENCIA ARTIFICIAL [online]. 18 (56), pp. 14–30.

9. Mucherino, A., Papajorgji, P.J. and Pardalos, P.M. (2009) k-Nearest Neighbor Classification. Data Mining in Agriculture [online]. pp. 83–106.