Predicting Student Performance using the APS based on a Conceptual Model

Khumo Tsagae
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
1929633@students.wits.ac.za

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

Kershree Padayachee
Science Teaching &
Learning Unit, Faculty of Science
The University of the Witwatersrand,
Johannesburg, South Africa
Kershree.Padayachee@wits.ac.za

***Abstract*—South African universities examine your overall APS score to see if you're qualified to study for the degree you're interested in. The purpose of this research is to apply data mining classification techniques to predict if a student would pass in record time given their Admission Point Score. This is accomplished by forecasting each student's performance class (low, medium, and high) on the synthetic dataset independently based on their overall grade for each subject before university entrance. There are three performance class aggregates examined (based on each subject weighting): low (0, 49), medium (50, 70), and high (70, 100), where each performance class contains 2000 partially imbalanced observations.**

The most accurate classification model is Logistic Regression, which has an Accuracy, Precision, and Recall of 99.95%, 100%, and 99.94% for the medium-performance class. The Support Vector Machine was the second-best model with an Accuracy, Precision, and Recall of 99.2%, 99.58%, and 99.46% for the medium-performance class. The Naive Bayes model has the lowest Accuracy, Precision, and Recall of 80.45%, 79.13%, and 83.86% for the high-performance class. For each of the four models, all the middle classes had the highest classification accuracy. According to the Information Gain table, the pre-school feature with the most information is English First Additional Language, which has the highest information gain (e) of 0.2024.

This research aims to assist South African Universities in classifying a student's university final year outcome (qualified or failed) based on their grade 12 results. It also indirectly examines the effect of lowering or raising the APS. This enables universities to examine how APS influences the number of record time graduates, in hopes of increasing the number.

***Index Terms*—Pre-school, Admission Point Score, Data Mining, Performance class**

I. Introduction

According to [11], the Admission Point Score (APS) is a metric developed by South African Universities to help students assess whether or not they are qualified to study for a certain undergraduate degree. It is a system used to inform students what grades they need to acquire to study in a particular field, [1]. According to the study, [11], the APS criteria is a weighted computation based on the symbols allocated to each subject in Grade 12, for example, taking a Computer Science degree requires at least 70% for Mathematics and 60% for English. As the APS is a determining metric on whether or not a student is accepted to study a particular course and it is directly measured using pre-school (grade 12) marks, is it enough to say that if a given student meets the university APS minimum requirements to study for a particular degree at a South African University, then they will pass their enrolled degree in record time? (note: Success in the context of this study is defined as completing a chosen degree in record time).

Although there is a large body of literature on higher education research on a variety of aspects, there is a scarcity of studies on establishing whether the APS is not just the right but also a sufficient metric to better predict student performance in South African Universities. According to the study, [10], Even while having the APS system intact, there is still a large number of problems faced by South African Universities such as failure rates, dropout rates, reduced attrition rates, and students' delayed progress. These problems have revealed themselves to be complicated, consistent, chronic, and insoluble challenges. For example, between 2008 and 2015, almost half of the students who completed the APS prerequisites for studying computer science failed to graduate in record time, as mentioned in the study [1]. The fact that the number of graduates has been declining for the past eight years, according to [1] highlights the need of evaluating the APS reliability.

This report investigates if meeting the APS standards is sufficient to say that a student will pass a chosen degree in record time. It is vital to investigate whether the APS is a trustworthy metric to determine whether South African Universities should continue to use the APS metric for university entry requirements or whether another system should be devised. This study is significant because it provides a deeper understanding of the APS and the link between pre-school performance and enrollment observations required to enable universities to predict and help at-risk students early so that they can graduate in record time in their chosen field. The APS impact is determined by predicting who is at risk of not making it. If three classs of people are given, one that barely made it to the admission standards, one that is average, and one that is far over the APS. This is accomplished by weighing the consequences of increasing the APS requirements.

Four Data mining methods namely Logistic Regression, Naive Bayes, Decision Tree, and Support Vector Machine were used on a synthetic dataset generated using a Bayesian network. The prediction of student performance is solely based on the pre-school and enrollment observations since we are considering the APS which is a weighted computation on individual subjects aggregates. The features considered are the pre-school, enrollment observations, and the target feature is Qualified.

The Literature Review (LR) which follow next, looks at the work of several authors who have tried to predict student performance. The LR determines the various traits that determine students' success using the APS. Firstly, it will provide the APS current state reliability. Secondly, it will conduct a thorough review of the literature of other writers and generalize how the literature supports the APS methodology. Finally, it concludes by discussing how the work of other writers influence and motivate this paper.

II. Related Work

Student performance is an essential metric for measuring students' and Universities long and short-term educational goals. By examining publications and linking the work of other writers by themes, the LR builds on the basic knowledge of forecasting student performance using APS. The themes related are the Data, Features selected, Methods used, and also the significance of different papers. The LR ends with a conclusion indicating the validity of the APS and other metrics using the reviewed literature.

A. APS current state reliability

Universities utilize the APS scoring system to advise students about courses they qualified to enroll in. The APS system assists South African Universities in dealing with a large number of applications. Some university courses need a higher APS score. According to [7] and [11], conventional measures of student ability (for example, the APS and the National Benchmark Test (NBT)), prior academic success (for example, high school GPA), and effort or motivation account for a considerable amount of the variance in student success. This brings up a significant point on the need to evaluate pre-school marks and the APS to assess student's performance.

The APS is used in conjunction with the National Benchmark Test (NBT). According to [7], the NBT assesses the writer's ability to apply knowledge of Academic Literacy, Quantitative Literacy, and Mathematics to post-secondary coursework requirements. If you do not fulfill the university's minimum entry criteria (APS), you can be denied admission to a particular course. To be accepted, you must submit a portfolio, audition, or attend an interview. When a particular student meets the APS to study for a particular course, the university assumes that the student has gained enough basic knowledge to study the course.

B. Data

Most of the papers' main aim is to determine and prove the factors affecting student performance in South African Universities utilizing enrolment features and biographical data. ([4], [7], [11], [2], and [3]) used historical real data from previous students to determine the features impacting students' success, while [10], [1] and, [12] used the Bayesian network to generate a synthetic dataset used in the study. [12] and [9] studies utilised both real and Bayesian network data.

According to [9], there is an increasing desire for more sophisticated methods of assessing valid educational data. This raises a question of which dataset is more reliable between primary or secondary datasets? The studies [9] and [1] consider future improvements on their work as using a real dataset.

The dataset utilized in this research is the same synthetic dataset used in the [8] study. The dataset has the potential of predicting student success using APS because it contained information about pre-school marks for subjects like Mathematics and English, and the NBT, which is a breakdown for APS. The prediction of student success is based on historical datasets, where this demonstrates that the prediction accuracy of these papers is dependent on the quality and quantity of recorded data.

Future research is needed to assess the consistency and accuracy of data utilized in these papers as mentioned by [1] and [3]. It also emphasizes the necessity of addressing reliability and validity when designing a study, looking for a secondary dataset, preparing your techniques, and putting up your findings, especially in quantitative research.

In the next section, feature selection, the relation to APS with the possibility of affecting student success will be discussed together with the feature selection employed in different papers.

C. Feature selection

As previously stated, the dataset used in a study must be reliable and valid. This is not sufficient if it does not take into account crucial features. All of the studied papers stress the importance of selecting the right features for the study. A possible feature that evidently relate to the APS as stated by the studies ([3], [12], [2], [11], and [7]) is the students grade 12 marks. Are there any other vital features to consider when applying for a course that should be compared to or utilized in conjunction with the APS system?

According to the studies [3] and [11], selecting essential features is not a simple operation because it necessitates the use of algorithms such as Information Gain Ranking (IGR), ChiSquared attribute evaluator, InfoGain attribute evaluator, OneR attribute evaluator, Symmetrical and Relief attribute evaluator, etc. Selecting important features should not be manual as it promotes biasness. The study [7] only focuses on determining important features that ensure student progress, where this shows the importance of choosing the right features to obtain good results.

Other features that indirectly contribute to the student success like the students' parent's occupation and salary as mentioned by the paper [12]. National Benchmark Test (NBT) as mentioned by [7] is used by the university to measure the students' ability to transfer understanding of Academic Literacy, Quantitative Literacy, and Mathematics to the demands of tertiary coursework. This raises a fact that the

APS score is not considered as the only determining factor on whether a student is accepted to enroll in a particular course or not, as mentioned by the study [1].

Following is the discussion of the models used in to write this paper. Possible enhancements to the research will be explored, as will their limits. The following section discusses data mining strategies for classification that have been utilized in the literature.

D. Methods

Most of the papers employed Data Mining Models of classification to binary classify a student into passed or failed using historical data to train the model. Classification is a supervised learning process in which the training of students' biographical and enrollment data is fed into a classifier that learns from classification rules. If the classifier is given test data later, it can predict the values if a given student passed or failed. According to [3], the classification entails predicting a specific consequence based on a given input, with the outcome being whether the student will complete their degree in record time or not, taking into account parameters such as the APS, NBT, pre-school marks, and so on.

Table 1 shows the benchmark accuracies for the models used in the literature. The accuracies are listed in order of decreasing accuracy.

Table 1 - Benchmark models and accuracies

Rank	Information Gain (e)	Accuracy (%)
1	Neural Network	84.6
2	Random Forest	83
3	Naive Bayes	80
4	Support Vector Machine	52

According to the study, [9], the Neural Network model and the Random Forest model had the highest accuracy of 84.6 percent and 83 percent respectively followed by the Naive Bayes model with an accuracy of 80 percent. The Support Vector Machine model has the lowest accuracy of 52 percent in both the studies [9] and [3].

As a result, data mining models, pre-school and enrollment features appear to have a great deal of predicting educational potential for us. For future improvements trials that include other variables such as psychological aspects that affect students, instructor motivating efforts, and e-learning tools available to students can be looked at as mentioned by the study [12].

E. Motivation of the study

The APS is a key determining metric of whether a student can enroll in a university course or not. As a result, it is crucial to determine if it is a good metric for an university entry requirements or whether it should be re-invented or utilized in conjunction with other entry requirements. The reliability of APS can be determined by the number of students that can graduate in record time in a given year. The proposed research focuses on assessing whether the APS can consistently produce a high proportion of students who pass in record time. It studies the effects of using pre-school marks on the number of students who graduate in record time.

III. Research Methodology

This study conducts a quantitative research methodology to fit a synthetic dataset generated using a learned Bayesian network on four data mining methods to classify a student into whether they managed to pass in record time or not. The dataset contained students' pre-school and enrollment mark aggregates, split into three classes (low $[0, 50)$, medium $[50, 70)$, and high $[70, 100]$). The low-class aggregate is less than 50%, the medium-class aggregate is between 50% and 70%, and the high-class aggregate is more than 70% for each subject mark respectively. The synthetic dataset is divided into three classes based on mark restrictions, with each class containing 2000 occurrences. The motivation for each class separation was to test the effect of altering the APS on the classification accuracy of each performance class.

The dataset utilized in this study is the same as that used in [8]. The synthetic data is generated using a learned Bayesian Network with forward sampling while assuming conditional dependence between the features. The details of the generation of the dataset are found in the study [8].

An overview discussion of the feature selection and contribution analysis, prediction models, and accuracy consideration follows. The data mining classification techniques classify a student if whether they qualified to pass in record time. The four data mining algorithms are the Logistic Regression, Decision tree, Naive Bayes Classifier, and Support Vector Machine. Consider the visual representation of the pipeline used to conduct the research:

The below-mentioned pipeline is used to conduct the research. The visual representation depicts the phases (stages) of this research paper. The subsections are organized in the same way as the pipeline.

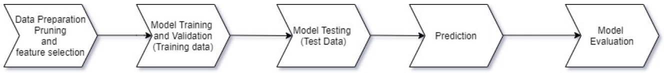
Data Preparation Pruning and feature selection
Model Training and Validation (Training data)
Model Testing (Test Data)
Prediction
Model Evaluation

Fig. 1. The pipeline used for this conducting research.

The initial stage is data preparation and feature selection, followed by model testing and validation via cross-validation. Model testing is used, followed by model evaluation utilizing confusion matrices and taking into account the accuracies, precision and recall of each model.

Data preprocessing and feature selection

The synthetic dataset used in this study initially had 41 features, spanning biographical, pre-school, and enrollment observations. Biographical information-containing features were removed, and only pre-school and enrollment marks-containing features were considered. The features were reduced to ten including the target label. The dataset initially had 14326 observations in total. The three features MathematicsMatricMajor, EnglishFirstLang, and EnglishFirstAdditional contained 37.2% of the missing data, where the features are spanned by the pre-school characteristic. The final data preparation step was removing the outliers using quantiles, and the null values were imputed using IterativeImputer. IterativeImputer replaces null values with Linear Regression, which is appropriate because students with the same grades in the same subjects are expected to perform similarly in the other subject.

The features that were considered for this study are displayed in the Figure 2 below, which is a visual representation of the conceptual framework (CF) used to conduct this study.

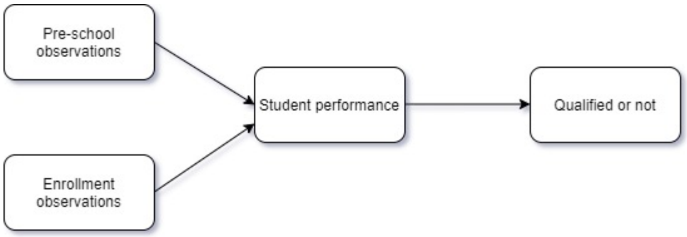

Fig. 2. The Conceptual framework used to conduct this study

The CF above summarises the 9 remaining features into two categories which are (i) Pre-school features = {Prior university enrollment mark aggregates for Life-Orientation, Mathematics, Mathematics Literacy, and the National Benchmark Test results for Mathematics and English which also contribute to university entry} and (ii) the Enrollment features = {Aggregate marks of a student in first, second, and third year of study }. The study by [8] explains the description of the ten remaining features (including target column).

The study [5] mentions that there is a variety of imputation methods available to replace nan values. The methods range from single to multiple value imputations. The imputation used in this study is called the Iterative-Imputer algorithm, which is a multiple value imputer that uses Linear Regression to replace nan values. The Iterative-Imputer algorithm does not introduce bias into multivariate estimations such as correlation and regression coefficients, and it beats single value imputation alternatives such as mean, mode, and median imputation. The single value imputation skews the correlation and variance of features thus it was not considered in this study.

Logistic Regression

In the presence of more than one explanatory variable, Logistic Regression is used to calculate the odds ratio. With the exception that the response variable is binomial, the approach is quite similar to multiple Linear Regression. The Logistic Regression model caters to outliers when classifying, unlike the Linear Regression model. The impact of each variable on the odds ratio of the observed event of interest is the end outcome {Y = Qualified}. The study [6] correctly states that "LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables." The Linear Regression model for *n* explanatory variables {X=X1, X2,..., Xn} is written as;

$$Pr(Qualified = 1) = 1(1 + e^{-}(\beta * X))$$

Where the above logistic function is called the Sigmoid function. Pr(Qualified=1) denotes the probability of qualifying in record time and the regression coefficients are given by β.

Decision Tree

A Decision Tree is a supervised learning approach that is non-parametric and may be used for both classification and regression problems. The aim is to train a model that predicts the value of a target variable (Qualified) using basic decision rules derived from data features. The algorithm traverses the pre-school and enrollment observations datasets efficiently while also defining a tree-like path to the expected outcomes (Qualified feature). The branching in a tree-like structure is based on feature values or control statements. If an internal node has a control statement (Life-Orientation $<$ 50), for example, the data points that satisfy this condition are on one side, while the rest are on the other.

The decision tree's structure begins at the root node and branches until it reaches the leaf nodes. An attribute selection measure defines the splitting, which is a heuristic for selecting the splitting criterion that is most capable of partitioning in a way that results in individual classes. Consider the attribute selection measure, Information Gain, which uses entropy to partition the features in a way that reduces randomness:

$$Info(D) = -\sum^{m} p_i * log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} |D_j D| * Info(D_j)$$

where p_i is the probability that an arbitrary tuple in a dataset D belongs to a class $C_i = |C_i, D||D|$. Info(D) is the mean entropy needed to determine a class for a data point in D. Then finally to calculate the Information gain:

$$Gain(A) = Info(D) - Info_A(D)$$

Since the dataset features were reduced from 40 to 10, the small dataset increased the likelihood that our Decision Tree model would not overfit the dataset by reducing correlated and redundant features. The Decision Tree was also chosen because it can handle both categorical and numerical features, and because of their non-parametric nature which makes no assumptions about the dataset.

Naive Bayes Classification

Naive Bayes is a classification technique that predicts the classification of incoming data based on historical data. It calculates the likelihood of an event occurring given the occurrence of another event. They enable us to forecast the likelihood of an event occurring based on the conditions we know about the occurrences in question.

The Naive Bayes classification technique assumes feature conditional independence, which indicates that the presence of one feature in a class is independent of the presence of any other feature. Even if the features are linked, they are nonetheless regarded as separate. This assumption gives rise to the method's name, which is why it is regarded as naive.

$$P(c|X) = \frac{P(X|c) * P(c)}{P(X)}$$

where: P(c|X): Represents the posterior probability of the target label c (Qualified), given the predictor features X.

P(c): Represents the prior probability of the target.

P(X|c): Represents the likelihood which is the probability of predictor X given the target label.

P(X): Represents the prior probability of the predictor.

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning approach for classification, regression, and anomaly detection. SVM work on the idea of identifying the best hyperplane given the other two planes that are closest to the support vectors and have marginal distances separating them from the best plane. The optimal plane splits the data into two categories based on the target variable. The SVM method tries to determine the distance between two object classes, assuming that the larger the distance, the more accurate the classification. The equation below shows the formulation of the soft SVM and it is followed by a diagram:

$$w^o, b^o = argmin\{\lambda * \|w\|^2 + \frac{1}{n} * \sum_{i=1} \zeta_i\}$$

such that

$$y_i(w * x_i + b) >= 1 - \zeta$$

where

$$\zeta_i$$

is the slack variable that measures the extent of violation of the constraint

$$y_i(w * x_i + b) >= 1 - \zeta$$

Consideration of accuracy

This section considers the metrics used to evaluate our model's accuracy. The synthetic dataset used encompass the response feature (Qualified feature), and the values of the feature are known; hence, a classification job is set-up. Confusion matrices, accuracy and entropy are used to quantify the correctness of our models during testing with cross validation.

Confusion matrices

The confusion matrix is a summary of classification problem prediction outcomes. The table below depicts a high-level overview of the confusion matrix:

Table 2 - Confusion Matrix

width=center

	P' (predicted)	N' (predicted)
P (actual)	TP	FN
N (actual)	FP	TN

(P and n) and (P and P') represent the model's actual and predicted results, respectively. A true positive (TP) is when the model accurately predicts the positive class. A true negative (TN), on the other hand, is a result in which the model correctly predicts the negative class. False-negative (FN) and False-positive (FP) denote the incorrect classifying of the students.

A. *Accuracy*

Accuracy is defined as the fraction of accurately predicted observations. It provides an answer to the question: What proportion of learners were correctly predicted by the model? High accuracy indicates that the model correctly predicted a substantial proportion of students. The metric formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. *Precision*

Precision according to the study, [9], is defined as the proportion of accurately predicted positive observations to the total number of expected positive observations. It provides an answer to the question: how many of the students who were predicted as qualified in record time truly qualified? High Precision indicates that students who qualified in record time are placed accurately. The metric formula is as follows:

$$Precision = \frac{TP}{TP + FP}$$

C. *Recall*

Recall according to the study, [9], is defined as the proportion of actual positives that our algorithm captures. It provides an answer to the question: How many students were predicted as qualified out of all those who are qualified to pass in record time? High recall indicates that a significant number of students who qualified in record time were correctly predicted. The metric formula is as follows:

$$Recall = \frac{TP}{TP + FN}$$

IV. Results and Discussion

A total of 4 Data Mining models were used to binary classify a student into qualified or not qualified using historical data to train the model. The four models include the Logistic Regression, Decision Tree, Support Vector Machine and Naive Bayes. The models classify each of the students from the three classes into qualified or not.

After running 10 fold Cross-Validation on the models using the three performance classes, Tables 4, 5, and 6 show the respective Accuracy, Precision, and Recall. Logistic

Regression is the most accurate classification model, with Accuracy, Precision, and Recall of 99.95%, 100%, and 99.94% for the medium-performance class. The Support Vector Machine was the second-best model in the medium-performance class, with Accuracy, Precision, and Recall of 99.2%, 99.58%, and 99.46%, respectively. For the high-performance class, the Naive Bayes model has the lowest Accuracy, Precision, and Recall of 80.45%, 79.13%, and 83.86%, respectively. Since the dataset was created using a synthetic dataset in which conditional dependence between features was assumed; this could explain why the Naive Bayes accuracy was low because it assumed that all features contributed independently to the target feature, which is not the case.

The fact that the Data Imputation was done with a method that is almost similar to Logistic Regression, Linear Regression, which may have influenced the high accuracy of the Logistic Regression. The Logistic Regression model was less prone to over-fitting because the dataset had a low dimension and a sufficient number of training samples for each sample class during the running of Cross Validation.

Confusion Matrices

Figures 3, 4, 5, and 6 depict the confusion matrices generated by a model using 10 fold cross-validation for each performance sub-class, each confusion matrix contains 2000 observations. The 10 confusion matrices were averaged after 10-fold cross-validation on each of the four models. Table 4 shows how the features rank in terms of Information Gain. Tables 4, 5, and 6 show the respective Accuracy, Precision, and Recall obtained by employing a model for each performance sub-class. The Information Gain quantifies a feature's relevance to the target variable.

Logistic Regression

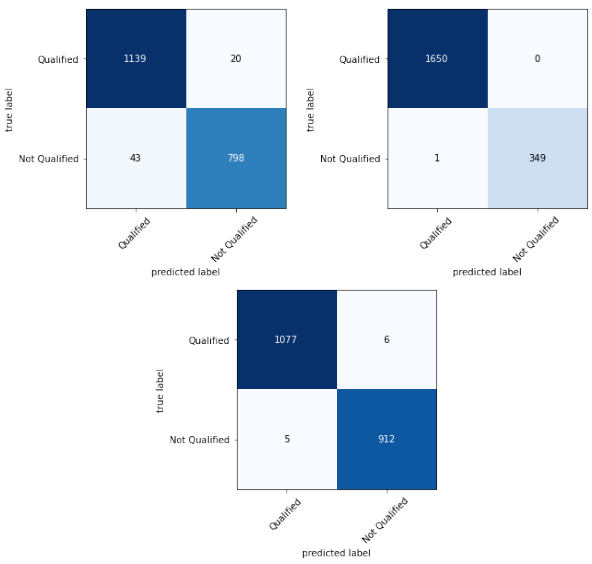

Fig. 3. Logistic Regression - low, medium, and high class respectively

Decision Tree

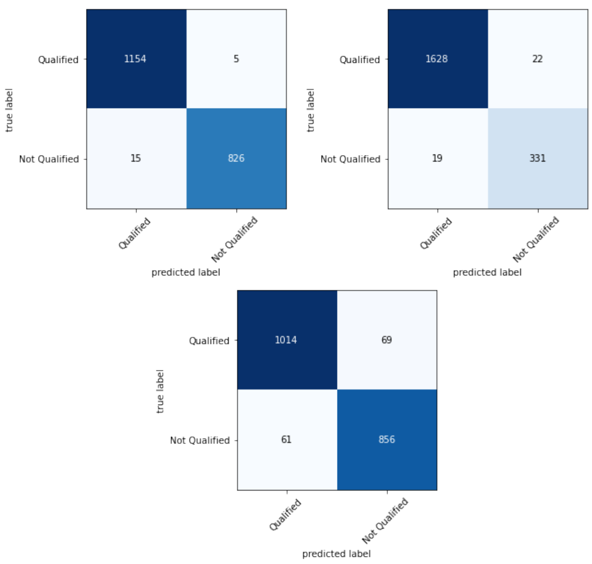

Fig. 4. Decision Tree - low, medium, and high class respectively

Naive Bayes

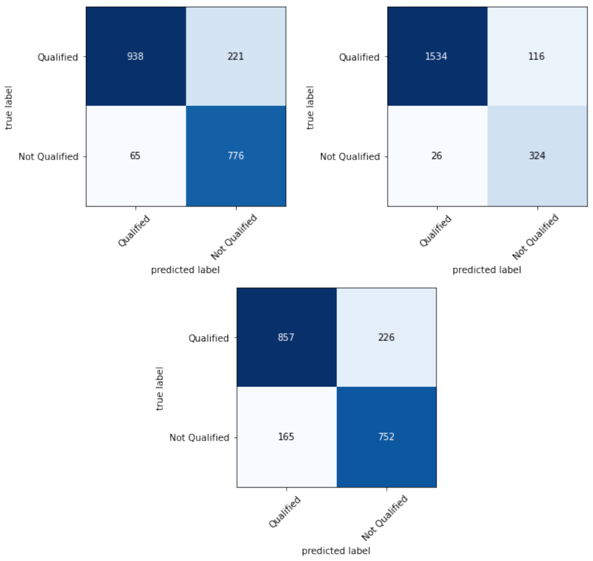

Fig. 5. Naive Bayes - low, medium, and high class respectively

Support Vector Machine

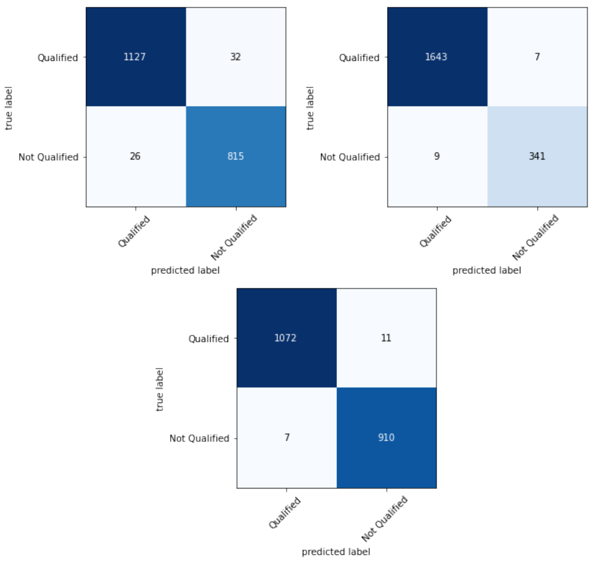

Fig. 6. Support Vector Machine - low, medium, and high class respectively

A. Information Gain - (IG)

The relevance of a subject weighted aggregate features to the Qualified variable is computed using information gain. It determines the statistical dependence or mutual information of two features. A high IG value (e) signifies low entropy and, as a result, reduced uncertainty between variables. The quantification of the relationship between the feature and the target variable is measured by using an IG value (e), where 0 $\leq e \leq 1$.

Table 3 - Information Gain

Rank	Information Gain (e)	Feature
1	0.448892	ThirdYearResults
2	0.290315	SecondYearResults
3	0.202408	EnglishFirstAdditional
4	0.193313	FirstYearResults
5	0.192335	NationalBenchMarkEnglish
6	0.174509	EnglishFirstLang
7	0.128610	NationalBenchMarkMaths
8	0.109442	LifeOrientation
9	0.093972	MathematicsMatricMajor

Table 4 - Model Accuracy

Model	Class Category	Class Accuracy (%)
Logistic Reg	Class 1	96.85
	Class 2	99.95
	Class 3	99.45
Decision Tree	Class 1	99.0
	Class 2	97.95
	Class 3	93.5
Naive Bayes	Class 1	85.74
	Class 2	92.97
	Class 3	80.45
SVM	Class 1	97.1
	Class 2	99.2
	Class 3	99.1

Table 5 - Model Precision

Model	Class Category	Class Precision (%)
Logistic Reg	Class 1	98.83
	Class 2	100
	Class 3	99.45
Decision Tree	Class 1	99.57
	Class 2	98.85
	Class 3	93.63
Naive Bayes	Class 1	80.93
	Class 2	92.97
	Class 3	79.13
SVM	Class 1	97.24
	Class 2	99.58
	Class 3	98.98

Table 6 - Model Recall

Model	Class Category	Class Precision (%)
Logistic Reg	Class 1	96.36
	Class 2	99.94
	Class 3	99.54
Decision Tree	Class 1	98.72
	Class 2	98.85
	Class 3	94.33
Naive Bayes	Class 1	93.52
	Class 2	98.33
	Class 3	83.86
SVM	Class 1	97.75
	Class 2	99.46
	Class 3	99.35

According to Table 3 (Information Gain) the features are divided into two groups which are pre-school and enrollment observations. The pre-school observations include EnglishFirstAdditional, Life-Orientation , MathematicsMatricMajor, NationalBenchMarkMaths, and NationalBenchMarkEnglish. The enrollment observations include FirstYearResults, SecondYearResults, and ThirdYearResults. The features that are more relevant to the target feature are SecondYearResults and ThirdYearResults. The APS is made up of pre-school features, and according to Table 4, the EnglishFirstAdditional and NationalBenchMarkEnglish features contain the most information gain and are thus the most important to consider when determining whether a student is qualified to pass in record time or not. According to the study, [11], the two features with the most Information Gain in determining student success were Mathematics and English.

V. Conclusion

The APS seems to be the only criterion for admission to a South African university. This necessitates a review of whether it is an appropriate metric for assessing student success.

This study focused solely on students' grades, which are subdivided into pre-school and university levels. The grades are categorized based on their mark ranges: low, medium, and high. This was done so that the impact of the APS may be studied individually for each class. No model scored lower than 80%, which suggests that the APS might be used to predict student performance.

According to the information gain table, third-year and second-year marks have a strong statistical dependence with the target variable where the features are part of enrollment observations. According to the findings of this study, English and Life-Orientation are the most correlated pre-school features when considering passing in record time hence them having the highest information gain of 0.202408 and 0.109442 respectively. Life-Orientation is not taken into account when calculating the APS, however since a synthetic dataset was employed, this could explain why it was considered in this study.

Model Overall Evaluation

All of the models have an Accuracy, Precision, and Recall that is greater than 80%, 79%, and 83% respectively. The achieved accuracies were higher than the benchmark as listed in Table 1. High accuracy was required in this study because misclassification can lead to a student being misled into pursuing a different degree, deviating them from their intended path. Accuracy only was not enough as mentioned by the study [9], hence Precision and Recall will further be investigated.

Predicting a student's performance is difficult because numerous elements influence whether or not a student will graduate on time. These characteristics include background, individual, pre-school, and enrollment observations, according to [1].

VI. Future Considerations

Future considerations could include (1) using an actual dataset rather than a synthetically created one, (2) incorporating all subjects into features, and (3) using pre-school marks to forecast every year outcome in a South African university so that we can be sure of a learner's potential from the beginning to the end. Predicting a student's future accurately is a difficult undertaking, as even one miscalculation can cost the university time and money.

References

[1] Tasneem Abed, Ritesh Ajoodha, and Ashwini Jadhav. A prediction model to improve student placement at a south african higher education institution. In *2020 International SAUPEC/RobMech/PRASA Conference*, pages 1–6. IEEE, 2020.

[2] Samy S Abu-Naser, Ihab S Zaqout, Mahmoud Abu Ghosh, Rasha R Atallah, and Eman Alajrami. Predicting student performance using artificial neural network: In the faculty of engineering and information technology. 2015.

[3] Ritesh Ajoodha, Ashwini Jadhav, and Shalini Dukhan. Forecasting learner attrition for student success at a south african university. In *Conference of the South African Institute of Computer Scientists and Information Technologists 2020*, pages 19–28, 2020.

[4] KG Bokana and DD Tewari. Determinants of student success at a south african university: An econometric analysis. *The Anthropologist*, 17(1):259–277, 2014.

[5] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17(1):1–10, 2017.

[6] Imran Kurt, Mevlut Ture, and A Turhan Kurum. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert systems with applications*, 34(1):366–374, 2008.

[7] Judith Goodness Khanyisa Mabunda, Ashwini Jadhav, and Ritesh Ajoodha. A review: Predicting student success at various levels of their learning journey in a science programme. 2020.

[8] Noluthando Mngadi. *A theoretical model to predict undergraduate learner attrition using background, individual, and schooling attributes*. PhD thesis, 2020.

[9] Ndou Ndiatenda. *THROUGH FORECASTING STUDENT SUCCESS IN HIGHER-EDUCATION*. PhD thesis, Faculty of Science, University of the Witwatersrand, Johannesburg, 2020.

[10] Ndiatenda Ndou, Ritesh Ajoodha, and Ashwini Jadhav. A case study to enhance student support initiatives through forecasting student success in higher-education. 2020.

[11] Thabo Ramaano, Ritesh Ajoodha, and Ashwini Jadhav. Different models relating prior computer experience with performance in first year computer science. 2021.

[12] VAMANAN Ramesh, P Parkavi, and K Ramar. Predicting student performance: a statistical and data mining approach. *International journal of computer applications*, 63(8), 2013.