Student Outcomes Prediction by used of Machine Learning

Mulweli Mudau
Computer Science
and Applied Mathematics
Faculty of Science
The University of the Witwatersrand
Johannesburg, South Africa
Email:1482422@students.wits.ac.za

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

Kershree Padayachee
Science Teaching and Learning Unit
Faculty of Science
The University of the Witwatersrand
Johannesburg, South Africa
Email:kershree.padayachee@wits.ac.za

***Abstract*—In this research, we utilize machine learning to predict student outcomes. The top performing machine learning algorithm with high accuracy may be used to create a system that assigns outcomes to students automatically. Six machine learning techniques are used to classify balanced (equally distributed on outcomes) synthetic data. Decision Trees, Random Forests, Supervised Vector Machines, Logistic Regression, K-Nearest Neighbor, and Naive Bayes are machine learning techniques employed. The performance of each method is determined by evaluating the models that have been built. Decision Trees performed quite well, with accuracy of 98 %, precision of 98 %, recall of 98 %, and F1-score of 98 %. Although Naive Bayes performed poorly, not all machine learning algorithms are accurate in forecasting outcomes. We can utilize decision trees to create a system that automatically assigns outcomes to students. Researchers that want to anticipate student outcomes might utilize Decision trees as one of their methods using varied data because it has a high rate of success in predicting outcomes in several studies, including this one.**

***Index Terms*—Student outcomes, Prediction,Higher Education, Machine Learning**

I. Introduction

Academic institutions review student grades at the conclusion of the academic year and decide what results are appropriate for specific students. The procedure must be carried out with caution. Different higher education institutions utilize different outcomes to determine which students should continue on to the next academic year, stay at the same level, or be expelled. There are significant distinctions between these outcomes, and staff members spend time attempting to make judgments for each student while keeping these distinctions in mind.Assigning academic results to students at the end of the year can be taxing on staff members since it necessitates sifting through large amounts of data to reach a conclusion. Machine learning can be effective in institutions when assigning outcomes since it can be used to analyse data and construct models that can predict data.

Researchers have shown a strong interest in academic research for student success, and data and traits that contribute to student success have been explored. Assigning student outcomes is an important aspect of student achievement; it determines whether or not a student will succeed. Machine learning has been used to predict student outcomes and whether or not students will achieve. [7] [8] [11] [13]. .

Six distinct machine learning algorithms are studied in this research to see how well they anticipate outcomes. The method with the best prediction performance can be applied in a system with similar data used in this study, to create a system that will automatically assign outcomes.

Data with five features and outcomes with 10 attributes is used in research for training and testing of different algorithms. Decision Trees, Random Forest, Supervised Vector Machines, Logistic Regression, K-Nearest Neighbor, and Naive Bayes were used to train the data in this study. Accuracy, Precision, Recall, F1-score, and Confusion Matrix are used to assess the models that have been trained.

When an examination board assigns outcomes to students, the most important factor in predicting outcomes is grades. Machine learning should have a high level of accuracy, according to studies that have used machine learning algorithms to assign outcomes. Accuracy above 80% is expected when dealing with assigning of outcomes.

A method that can be used to build a system that assign outcomes using data of same format is determined from six machine learning Decision Trees, Random Forest, Supervised Vector Machines, Logistic Regression, K-Nearest Neighbour and Naive Bayes.There are methods that did not perform well, such methods researchers might want to exclude for future researches.

This paper begins with introduction section I, following will be related work literature review section II ,section II: methodology how one can obtain results from data to, methods used and techniques used for machine learning evaluation, section IV: results obtained in research, section V:discussion and conclusion.

II. Related work

Few studies have looked into how outcomes can be given automatically using a machine learning system.Since different institutions use distinct outcome classes, the focus of automatic outcome assignment has not been on institutions as a whole. MNRN: when a student is expelled due to low performance and must appeal to be re-admitted, for example. If the student continues to perform poorly, MBR:exclusion will be waived. RET: student is not permitted to continue to the next academic year and must repeat the current academic year; PCD: student is permitted to continue to the next academic year. [7].

Data-sets in researches that have been done to predict outcomes or predicting success have used different features, some have done research on which features contribute more to prediction of outcomes. The most important feature is Grades / marks of student in predicting outcome. Information gain has been used to gain information on which features contribute predicting outcomes and predicting success. This features include Year of Study,Course Code,Final Mark,Final Grade and Grade ID or Encrypted Student No, Gender, Nationality, ProgramCode and Behavioural [8] [13] [14] My research uses Year of Study,Course Code,Final Mark,Final Grade and Encrypted Student No to predict outcomes.

Researches related to Educational studies involving prediction of outcomes and prediction of student performance have show that use of machine learning techniques like Logistic Regression, Naive Bayes, Supervised Vector Machines, K-Nearest Neighbour, Decision Trees, Random Forest and Multilayer Perception, [12] [10] [8]. Random Forest which is a combination of decision trees have performed with high accuracy and precision in predicting student outcomes [7]. In this research Decision Trees, Random Forest, Supervised Vector Machines, Logistic Regression, K-Nearest Neighbour and Naive Bayes.

TABLE I contains related articles to this research with features, data, models and accuracy.

Research papers			
Author	Data	Features	Models
Jettiniel	WITS University	Student results	Supporting vector machine, Random forest, Na¨ıve Bayes, Classification using Regression and Multilayer Perception
Prince	High-Education Research-Intensive Institution (HEIs)	Computer Science degree, from first year results	J48, Na¨ıve Bayes, Support vector machines, Multilayer Perceptron, Logistic Regression and the k* w
Gcobisile	WITS University	student marks 2010-2017	Na¨ıve Bayes, Support Vector Machine and Decision Tree
Harmony	WITS University	Biographic data and enrollment information	mining methods

TABLE I: Related articles

III. Methodology

We are looking for a machine learning system that can assign outcomes to students in this study [8]. Different machine learning models are developed using data with five features and outputs with ten types of outcomes. Six alternative machine learning algorithms are trained and their predictions are tested. The accuracy, precision, recall, f1-score, and confusion matrix are used to assess the approaches' performance.

A. Data and Features

Synthetic data, data generated which reflects data from University of the Witwatersrand(WITS) is used when conduction research. This data contain five features: Year of Study,Course Code,Final Mark,Final Grade and Encrypted Student No.and a feature of outcomes that the is being prediction. There are ten attributes outcomes ,FTC-Failed to complete requirements for qualification, M1C-Renewal of registration permitted by Wits Readmission Committee-1 with conditions,M2E-Permission to renew registration refused by Wits Readmission Committee-2, MBR-Exclusion waived return to same year of study, MBZ-Failed minimum requirements-Needs permission to re-register in same faculty,NCD-Credit bearing at discretion of home institution, PCD-Permitted to proceed,Q-Completed all requirements for qualification, RCC-Candidature continues and RET-Must return to complete requirements for year of study.

Data used in the research was imbalanced in outcomes , outcomes range was to high with FTCS 247 times whilst at highest we had PCD at 9179 and others were between this in an uneven distribution.Machine learning methods mostly are build to classify data that is balanced [6]. Trained imbalanced data may lead to the predictive models favouring the majority class while making prediction. Having balanced data gives priority to each class.Balancing techniques SMOTE and Oversampling were applied to the data in order to have data that is equally distributed in different outcomes given to students [4] [5]. The balanced data was used to in

classification.

B. Methods

1) Naive Bayes classifier: It is a classification technique based on Bayes Theorem with an assumption of independence among predictors.In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [Romero and Ventura 2007]. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to.outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c—x) from P(c), P(x) and P(x—c). Look at the equation below:

• P(c—x) is the posterior probability of class (c,target) given predictor (x,attributes). • P(x) is the prior probability of class.

$$P(x_i|y, x_1, ..., x_n) = \frac{P(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)}$$

Using the naive independence assumption that

$$P(x_i|y, x_1, ..., x_i - 1, x_i + 1, ..., x_n) = P(x_i|y)$$

for all i, this relationship is simplified to

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)}$$

Since

$$P(x_1, ..., x_n)$$

is constant given the input,we can use the following classification rule:

$$P(y|x_1, ..., x_n)\alpha P(y)\prod_{i=1}^{n} P(x_i|y)$$

$$\widehat{y} = \underset{y}{\operatorname{argmax}} P(y)\prod_{i=1}^{n} P(x_i|y)$$

Figure 3.2: Bayes equation

2) Decision Trees: This a flowchart like structure with tree layout with internal node, branches and leafs. Where internal node is our test on each feature , branches representing outcomes of test and each leaf node represents class label.

3) Logistic Regression: Logistic regression is a type of regression method predicts new value for data we given. Probability is used directly using Logit transform. [9]This method is similar to linear regression, they differ because logistic regression is used for classification. When data features focused on include biographic data and enrollment data the method yield one of the best accuracy in student result prediction.

4) Support Vector Machine: Support vector machines are predictive methods associated with algorithms which analyze data for classification and regression analysis. This are powerful predictive methods based on statistical learning of frameworks. Support Vector Machine is a learning algorithm supervised, it determines hyper-plane that separates two classes by greatest margin between them.Data is casted to higher dimensional plane on which it can be divided. The methods finds optimal hyper-plane that linearly divides data. [10]

5) Random Forest: Random forest is a machine learning algorithm that creates a forest, which is a set of decision trees, using the bagging process. The approach is simple to use and is based on supervised learning. To find an accurate and stable estimate, multiple decision trees are created and combined.

6) K-Nearest Neighbour: K-Nearest Neighbour(KNN) is a supervised machine learning that solves classification and regression problems.The method finds distance between examples in data-set and query by taking specific number of K examples that are closest to the query, then selects the frequent label or takes averages of labels if its regression [3].

C. Evaluation methods

• Accuracy- how many predictions got right.

• Precision-Out of predictions how many are correct.Precision evaluation method tells the correctly predicted cases that turned to be true. We can tell if the model that is being used is reliable or not. Precision can be calculated as follows:

TP: True positives

FP: False positives

FN: False negatives

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

• Recall-Correct predictions out of the truth.

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

•F1-score: Overall of recall and precision.F1-score is the harmonic mean of recall and precision.

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

• Confusion Matrix -Confusion matrix is a classification model assessment approach that uses a N x N matrix, with N being the number of needed classes. The machine learning model's projected values are compared to the

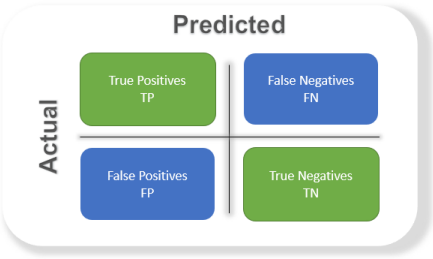

Fig. 1: Confusion Matrix

actual target values. This enables us to comprehend the categorization method's performance and faults.Fig 1 shows a confused 2 x 2 matrix for a binary classification task.

IV. Results and Discussion

Models	Accuracy %	Precision %	Recall %	F1-score %
Decision Trees	98	98	98	98
Random Forest	87	87	87	87
K-Nearest Neighbour	82	81	82	81
Logistic regression	59	21	25	23
Support Vector Machines(SVMs)	58	84	22	25
Naive Bayes	43	35	17	15

TABLE II: Classification Report.

The experiment was carried out using six machine learning models, which are presented here. Year of Study, Course Code, Final Mark, Final Grade, and Encrypted Student No. are used to predict ten outcomes. The 10 outcomes are allocated to integers ranging from 0 to 9. M1C:1, M2E:2, MBR:3, MBZ:4, NCD:5, PCD:6, Q:7, RCC: 8, and RET:9. FTC:0, M1C:1, M2E:2, MBR:3, MBZ:4, NCD:5, PCD:6, Q:7, RCC: 8, and RET:9. The Confusion Matrix and Classification Report were used to examine how well each method's prediction performed in terms of predicting outcomes. For our testing data, the confusion matrix compares how each attribute/outcome is predicted.Classification report comprises of precision which tells us out of all our predictions which ones are correct for each attribute, recall which tells us how many predictions are correct compared to the truth for each attribute and f1-score which is overall performance of precision of each attribute and accuracy which gives us overall correct prediction.

col_0	0	2	4	6	7	9
Outcome						
0	1003	0	0	802	31	0
1	5	0	0	9	38	0
2	7	0	0	55	21	0
3	3	0	0	242	191	0
4	45	0	1	87	59	0
5	7	0	0	71	2	0
6	24	1	0	896	915	0
7	36	0	0	420	1065	0
8	56	0	0	120	42	0
9	32	0	0	299	381	2

Listing 1: Confusion Matrix

A. Naive Bayes

The results obtained when implementing Naive Bayes are as follows and can be read from classification report in IIand the confusion matrix figure 4.2. The overall precision for all attributes is 35 % , overall recall 17% , f1-score 15% and accuracy is 43% .

The confusion matrix reveals that when making predictions, this approach ignored qualities 1, 3, 5, and 8. One of Naive Bayes' drawbacks is that categorical variables whose category in the test data set was not available in the training dataset. Also, we can observe from the confusion matrix that the outcomes that are appropriately not understood as other outcomes are less than those that are accurately predicted.

B. Support vector machines (SVMs)

col_0	0	1	2	3	4	6	7	8	9
Outcome									
0	2	0	0	0	0	40	0	0	0
1	0	7	0	0	0	42	0	0	0
2	0	0	1	0	0	85	0	0	1
3	0	0	0	77	0	374	1	0	2
4	0	0	0	0	17	173	0	0	0
5	0	0	0	0	0	70	0	0	0
6	0	0	0	0	0	1837	0	0	0
7	0	0	0	0	0	570	940	0	0
8	0	0	0	0	0	227	0	1	0
9	0	0	0	0	0	606	0	0	109

Listing 2: Confusion Matrix

Support Vector machines(SVMs) performed better than Naive Bayes. The overall precision of the method is very high at 84 % with 7 of outcomes being having 100% precision. One outcome This method disregarded attributes 1, 3, 5 and 8 when doing prediction. 'NCD':5 is disregarded when doing prediction. Overall recall is at 22% , the prediction over the truth percentage is low for most outcomes. F1-score is at 25% and accuracy at 58%.

Predicted outcomes are false for most outcomes , only one outcome got predicted correctly for all students assigned to the outcome. The rest of outcomes are predicted to be other outcomes for most students.

C. Logistic Regression

This model disregard 4 outcomes which leads to accuracy not being reliable. The accuracy is at 59% whilst overall precision is at 21% , recall 25% and f1-score 3 % . Using the data had this method can not be used in a system to assign outcomes to students.

The individual prediction of each outcome not satisfactory. There are more outcomes being predicted thus assigning outcome to students who did not receive that outcome as in the testing data. The confusion matrix in Listing 3 shows this for outcome 3 up to 9 being incorrectly predicted for more students in the training dataset.

col_0	0	3	6	7	9
Outcome					
0	1605	25	206	0	0
1	10	19	6	17	0
2	38	28	16	1	0
3	128	96	93	118	1
4	106	48	21	17	0
5	18	8	53	1	0
6	53	45	1363	374	1
7	66	3	408	1044	0
8	101	0	117	0	0
9	114	111	264	225	0

Listing 3: Confusion Matrix

D. k-Neighbour

The accuracy of this method is 82% which is reliable when looking at overall precision is at 81% , recall at 82% and f1-score at 81%.

Individual outcomes are predicted correctly in this model which can be observed in the confusion matrix in Listing 4. We are looking for model which performs at almost 100% accuracy so that staff members don't have to go through every results in order to assign outcomes according to rules.This model still not reliable to be used using data I had.

col_0	FTC	M1C	M2E	MBR	MBZ	NCD	PCD	Q	RCC	RET
Outcome										
FTC	1669	8	18	7	21	7	1	4	86	15
M1C	8	1725	17	36	14	10	4	2	3	17
M2E	7	16	1652	36	50	23	9	4	5	33
MBR	7	43	77	1466	77	42	42	23	6	53
MBZ	35	19	37	80	1587	22	11	7	14	24
NCD	7	21	24	55	33	1591	24	34	12	35
PCD	14	34	49	138	47	142	1016	165	29	201
Q	26	32	9	46	20	59	126	1443	29	46
RCC	101	16	10	14	21	8	16	12	1626	12
RET	32	44	59	116	73	91	130	53	51	1187

Listing 4: Confusion Matrix

E. Random Forest

The overall prediction is at 87% for Random forest. The accuracy is 87% , precision is 87% , f1-score is also 87% and recall.

The confusion matrix Listing 5 for random forest gives us a picture on how individual outcomes are being predicted correctly for each outcome. Very few students are assigned to wrong outcome in our testing dataset. The model is better than the last 4 above.

col_0	FTC	M1C	M2E	MBR	MBZ	NCD	PCD	Q	RCC	RET
Outcome										
FTC	1717	3	3	4	2	5	3	2	90	7
M1C	2	1792	8	6	4	7	7	0	1	9
M2E	2	8	1704	39	25	25	14	1	1	16
MBR	7	26	70	1454	75	39	59	14	2	90
MBZ	9	26	42	77	1607	22	19	2	5	27
NCD	3	7	12	22	24	1724	9	12	8	15
PCD	8	11	23	62	16	19	1515	5	16	160
Q	6	6	4	5	3	14	19	1740	23	16
RCC	122	2	3	2	3	9	6	18	1655	16
RET	41	33	41	160	48	45	272	43	61	1092

Listing 5: Confusion Matrix

F. Decision Tree

Decision tree classifier proved to be the best classification method in this research. It has an accuracy of 98% , precision 98% ,recall 98% and f1-score 98%.

Predictions are accurate almost at 100% for each outcome. Students in training dataset are correctly assigned to outcome. Only four outcomes have students being assigned to wrong outcome which is occurring at a very low percentage.

col_0	0	1	2	3	4	5	6	7	8	9
Outcome										
0	1836	0	0	0	0	0	0	0	0	0
1	0	1836	0	0	0	0	0	0	0	0
2	0	0	1835	0	0	0	0	0	0	0
3	1	0	0	1816	2	1	4	1	0	11
4	0	0	0	0	1834	0	0	0	0	2
5	0	0	0	0	0	1836	0	0	0	0
6	2	3	11	35	13	3	1651	1	3	113
7	3	0	0	14	1	0	1	1796	3	18
8	0	0	0	0	0	0	0	0	1836	0
9	3	3	0	30	11	0	56	16	12	1705

Listing 6: Confusion Matrix

G. Analysis

Decision tree is definitely the best performing model out of the six approaches used in this study,with an accuracy of 98% , precision 98% ,recall 98% and f1-score 98%.Decision tree is able to lay out the problem clearly so that all options can be challenged.The next best performing model is Random Forest at accuracy is 87% , precision is 87% , f1-score is also 87% and recall.Random Forest fails to find significance of each variable, we can observe in Listing 5 that predictions are made incorrectly. K-Nearest Neighbour has accuracy 82%, precision is at 81% , recall at 82% and f1-score at 81%. The performance of K-Nearest Neighbour is fairly high but this method does not operate well when dealing with high dimension data, calculating distance in each dimension is difficult [1].

For Logistic Regression accuracy is at 59% whilst precision is at 21% , recall 25% and f1-score 3 % . SVMs has

Overall is at 22% ,F1-score is at 25% , accuracy at 58% and Precision of 84%. To obtain more accurate results using SVMs good kernel must be found, this can be difficult, in my research I could not find better kernel to improve predicting performance.Logistic regression assumes linearly relationship between dependent and independent variable, this is not always the case.The method with poorest performance is Naive Bayes with precision for all attributes is 35 % , recall 17% , f1-score 15% and accuracy is 43%.Naive Bayes makes assumption that features are independent when making classification which is not the case in our data this results in poor performance. The method also experience zero frequency problem if data set in testing set is not in training data it assigns it to zero probability thus making wrong prediction [2]. Such problem can be cause that leads to Listing 1.

V. Conclusions

The study determines a method/model that can be used to create a system for allocating student outcomes. Naive Bayes, SVMs, Logistic regression, K-Neighbour, Random Forest, and Decision Trees are among the machine learning algorithms employed. In order to forecast outcomes, synthetic data is used. There are five features in the data: Year of Study is used to determine which students are in which year of study. The course code is used to determine which students are enrolled in the same courses. The feature included a final grade that indicated whether or not the student passed the module, as well as a classifier that identified all of the modules the student was studying. The crucial feature of information gained by models is encrypted Student No and Final Mark.

The classifiers are used to predicted data in the training set and Decision Tree performed best with recall of 98%. The second best was random forest with recall of 87%. Then it was K-Neighbour which performed at 82% recall. Logistic regression performance followed at 25% recall . Support Vector Machines(SVMs) had recall of 22% second before the model that performed lowest. Naive Bayes performed at 17% recall which was the worst of the six methods implemented on this research.

In summary Decision Tree performed best compared to Naive Bayes, SVMs, Logistic regression , K-Neighbour and Random Forest. Machine learning algorithms proved to be great at assigning outcome on this research. Given Year of Study, Course Code, Final Grade , Encrypted Student No and student Final Mark a system made of Decision Tree can be used to assign student outcomes. In order to lessen burden that examination board had when assigning outcomes this research contributes by finding that machine learning can be used to assign outcomes and it would be efficient in studying the boundaries between outcomes unlike the board members. From this research a conclusion can be draw that machine learning can be used to assign student outcomes, in this case if have data similar to this research one can use Decision Tree to build the system that will assign outcomes.In future one can predict student outcomes making restriction on year of study in which student is on since some outcomes only apply to student at certain academic year of study.

References

[1] Advantages and disadvantages of knn algorithm in machine learning.

[2] Decision tree analysis: Choosing by projecting ”expected outcomes”.

[3] Machine learning basics with the k-nearest neighbors algorithm.

[4] Random oversampling and undersampling for imbalanced classification.

[5] Smote for imbalanced classification with python.

[6] Why balancing your data set is important?

[7] J. Chepiri. . automatic labelling of student results as pcd, ret, mbr and mrnm. *Unpuplished, Computer Science and Applied Mathematics. The University of the Witwatersrand, Johannesburg.*, 2017.

[8] R. A. Eluwumi Buraimoh and K. Padayachee. Predicting student success at various levels of their learning journey in a science programme. *The International IOT, Electronics and Mechatronics Conference (IEMTRONICS 2021) in association with IEEE Vancouver section, Vancouver, Canada.*, 2021.

[9] A. J. n. R. A. Lonia Masangu1. Predicting student academic performance using data mining techniques. *Advances in Science, Technology and Engineering Systems Journal*, 6(1):153–163, 2021.

[10] G. Matafeni. Using big data analytics to predict learner attrition based on first year marks at a south african university. *Advances in Science, Technology and Engineering Systems Journal. Special Issue on Multidisciplinary Sciences and Engineering.*, 5, 2020.

[11] H. Mncube. . predicting student performance using first year marks. *—Unpublished, School of Computer Science and Applied Mathematics.*, 2020.

[12] D. R. A. Nastassja Jacqueline Philippou and D. A. Jadhav. Using machine learning techniques and matric grades to predict the success of first year university students. *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2020.

[13] P. Ngema and D. R. Adjhooda. Predicting student performance using enrolment figures. school of computer science and applied mathematics. *Unpublished, School of Computer Science and Applied Mathematics. The University of the Witwatersrand, Johannesburg.*, 2019.

[14] S. K. Yadav and Saurabh. Data mining: A prediction for performance improvement of engineering students using classification. *Technology Journal (WCSIT) ISSN: 2221-0741*, 2(2):51–56, 2012.