Predicting Student Performance Using Enrollment Figures and Biographical Information

Harmony Mncube(1272371)

***Abstract*—In order to predict students' performance in future academics, an educational institution needs a rough earlier information from enrolled students. The ability to anticipate the performance of a student could be beneficial to the institution, a number of different ways shows a method of dealing with an auto-labeling system that looks at the student and ascribes a result code. It will help to determine the anticipated future outcome for a specific student. This enables the university to pinpoint promising students as well as to give the university a chance to pay attention and improve those that are likely to receive lower grades or are at risk of potentially failing to complete the current year under investigation.**

In this paper, I designed models that can anticipate students' achievement based on their previous performances using data mining methods of Classification. The results from the analysis shows that various predictive models (i.e.decision trees) achieve a sustainable accuracy that implies the possibility of better presumptions of a student's success in his or her career at university. I further sideline with the prediction of students outcome based on only just biographic data and student outcome.

***Index Terms*—Naive Bayes, Decision trees ,Linear regression, Feed Foward Neural Networks, python3, matlab**

I. Introduction

South African institutions are seriously concerned with the low pass rate of students in science fields notwithstanding the strategies introduced to deal with this issue [1]. Less selective universities are using predictive analytics to improve student outcomes, especially among those who are struggling academically. The use of predictive analytics involves "the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data", with the objective of providing the best assessment for what will happen in the future.
Its use enables universities to swoop in and grant support for students at-risk of dropping out before things take a turn for the worse, thus improving student outcome. The expanding access toward the South African advanced education system has brought about the affirmation of a substantial companion of students from impeded foundations [1]. While this is a positive example, an investigative study on a partner of school students in South Africa uncovered that the most greatest attrition rate occurred close to the completion of the essential year of study (29% of first year students) [1]. Only 30% of the first-time students had graduated following a five-year term [1]

The authors, [1] also extend to point out a high influence from biographical associations and previous performance among other external factors beside just looking at academic performance. Authors [2] talks about the influence of family socioeconomic status on the academic performance of student. Family income plays a somewhat role in the students performance in his/her academic career. Some socioeconomic factors don't really have a direct influences on students' educational outcomes [2].

Therefore, educational institution needs a prior knowledge learning of enrolled students to anticipate their performance in future academics [1]. This helps them to recognise promising students and furthermore gives them a chance to focus on and improve the individuals who might presumably get lower grades despite the inadequate level of education given at especially disadvantaged schools whereby majority of the learners subsequently fall under.

In order to continue with this research, I used data provided by WITS university, the dataset contains information about students, such as race, biographical data and marks from the previous year of study which will be analyzed. multiple various models ,some which are based in 1986 by [3] on Hunts algorithm(Decision Trees e.g CART) will be used, and also Naive Bayes, Feed forward Artificial Neural Neural Network and with regression. By applying these algorithms to this data, in future examinations we anticipated the aggregate and individual reliability of any given students.

This research will try to identify at-risk student using previous enrolment observations through various data mining operations to analyze and evaluate academic data from undergraduates and also improving the quality of the university system. The main contributions of this paper are: (a) Indication of the influence of enrollment figures of student; (b) Student performance with respect to academic aggregate characteristics over the period 2015 to 2018; and finally (c) Compare the results outcome form all the algorithms used.

These classification model could be used by higher management to improve the course overall result according to the knowledge extracted [4]. Such understanding could also be used for a broader understanding of student enrollment trends throughout the course being examined, as well as the faculty and managerial decision-maker to use the necessary measures to provide extra pre-requisites and academic counseling. Using such knowledge, the management system can improve its policies, improve its techniques improving the quality of that same management system [4]. Nonetheless, several other classification methods can also be applied to test the most suitable method that suit the structure of the student data and give a better classification accuracy.

My research provides a process used to design and create a prediction model artefact that predicts academic performance of students. It stands out for operating longer than others, in addition to its success in improving outcomes "with low-income, first-generation and minority students – groups that have the hardest time succeeding in higher education".

While it may seem premature to dash a students' dream of completing their chosen major before they run into trouble with their performance, some universities may see this as an effective way to tackle university dropout rates.

These findings will contribute to the knowledge discovery in students performance also in data mining. I will uncover some algorithms and methods, this will incorporate problems faced in data manipulation that usually influence the academic performance prediction model in order to have a successful design, implementation including its adoption. Prediction models to foresee students outcome will benefit institutions more with extensive analyses of factors associated with enrollment outcomes.

II. Related Work

A. Introduction

Increasing student persistence in all educational institutions is a long-term goal. Some authors highlight the importance of attrition on students' academics. After all different student apply different strategies in getting through their academic program. From [5] research, the author implied that traditionally, logistic regression and discriminant analysis are frequently used in retention studies of students academic success or failures. However, logistic regression is bias in large data and is unreliable [5].

Good data results in good results otherwise bad data results in inconsiderable decision making thus using supervised machine learning in labelled data as suggested by the author [6] that he used reinforcement learning and the measure of how well the system operates to help with the validation of students performance. Some aspects for data mining are in graduation, and academic performance on web-based education etc [5]. We may rely on student participation on traceable systems to allow grouping of students based on similar trends but supervised machine learning techniques are not that easy in constructing intricate student models [7]. In this case, features of students' model, the students' behavior, and the background knowledge were considered but however, we can not rely only on background characteristics alone [5], because there are other factors that may influence student progress.

Research on educational studies consists in fact of various sub-field [8]. This approach describes a significant proportion of early learning work that has created ego-improvement systems for critical thinking. Learning in the form of building rules is seen as the prolonged development of organized information in a different strategy.

B. Background

1) Introduction: This paper explores both biographical data and enrollment observations, that may contribute to the success of the student. Training examples of the form $f(x_i, y_i)$ for x $\in$ R^n where y = f(x) and x values are real-values such as age.

The y values are taken from a set of classification classes e.g y =1,..., K where $y_i \in R$ represents the outcome for a student (0=PCD, 1=RET, etc) in the case of classification. x_{ij} will represent the j-th feature of x_i's. The model consist of numerical value for ease use of data set, string are converted to numerical values and stored on a separate dataset (the important background information relating to the student is usually in string format ,e.g. race). The error of the model is monitored closely to see if we not drifting away from the expected results. Then the constructed student model classifies and predict outcome of students based on both enrollment figures,past test and or other results including background computed information . From the author [9], ID3 is used in computing valid outputs based on the data structures per attributes which is the same method that we will inherit here.

2) Domain Complexity: Complexities of domain and domain tasks are usually alluded because the concept of learning or classification task in the sense of attacking problems requires the classification ability. Complex linear student models are easily solved by PIXIE compared to ASSERT which models students' classification represented as vectors. Therefore, viewing these as a spectrum or possible multidimensional for instance with mathematics and programming can allow easy manipulation [10].

3) The Student Model: The student's main statistical description of a specific field by [10] somehow accounts for student behaviour. By being qualitative, we simple mean that it is either numerical (information in quantities; that is, information that can be measured and written down with numbers). This model can only account for computational utility rather than in cognitive fidelity [10].

4) Levels of error analysis: Relationship between the actual and desired behaviours are determined , these discrepancies behavior or behavioral-level errors are named after incorrect behaviour based on their importance when dealing with simple integer behavior. It then becomes non-trivial when we trade our interests to less valuable task [10] (more complex behaviours like programs) resulting in the significance of this knowledge level meaning that behavioral error can also be due to inconsistency or and insufficient knowledge [8], [10]. Therefore ,all information is important because it gives a statical picture [8] .

Usually the student model is concerned with the error at learning level, this allows the model to know what it is dealing with for better prediction accuracy. Since our model is to predict students' outcome, therefore, having a strong model can bring about a better accuracy in predictions.

5) The Student Model Construction: There are two approach that seem rather considerable compared to other methods, I looked at student academic progress knowledge to transform the student behaviour so that I can quantify the relationship between the student behaviour and the problem given if they have some common relationship to classify the student progress so as to create a strong model. The second approach based on the author [10] inherited was to synthesise elements from the fuculty registrations to devise an analytic approach, while taking a close look at system that construct their student models from multiple behaviour. All of these methods reduce records (removing redundant and irrelevant features) to allow data mining algorithms to operate effectively as some random noise can be influenced by the examples of training.

The background knowledge features is a set of unrelated values. These values are calculated based on information gathered about a specific student.

6) Logic Based Algorithms and Related Work: The predictions institutional research in higher education is a focus point. However, predicting various cases such as student dropout, graduation duration is not an easy process [11]. Neural networks have been used previouslu by various researchers for predictions of students' results. The author [11] presents a neural network-based decision support system that identifies students who are "at-risk" of not possible graduating in record time, completing the degree overall meaning that they dropped out, repeating a specific coarse and other challenges that students face. The author [11] used regression and path analysis but these contributed a little to understanding of student retention. Therefore, [11] then turned to use neural networks to learn the student models so that they can be able to predict differences course in these various cases such as time taken to complete a degree, the number of student who returned to the same grade etc. The results suggested that level of complexity of the data used and the outcome predicted may largely contribute into selecting the right attributes but inferential statistics with little collinearity among variables may not be the best use of decision tree or neural net methods

[5] used some important demographic converted values and assignment marks to anticipate students' academic performance using decision trees and logistics regression algorithms and other algorithms that are not relevant here. His decision tree classified students based on features with high information gain at root node. NP-complete was then used as their appropriate way of constructing binary decision trees optimally for final output/prediction. [6] emphasise starting from the root node, with the feature that describe our data better using information gain, with the aid of myopic based on [8]'s method to measure estimations of each attribute independently proceeding down to the leaves/decision. The same procedure is adopted by [6] creating branches for each partition of the divided data, until subsets are of the same class. The results were not as accurate as expected because as few methods for comparison were used

[5], [12] used ID3 to incorporate method for building the tree iteratively, then Bayesian was applied to determine the probability to belong to a specific class. However, excluded variables reduced the accuracy of naive Bayes classifier for predicting students' performance [12]. [12] analysed students based on decision trees to classify them under low, medium and high-risk students, this method will be used in our final system we built.

At risk students research on Mathematical Sciences was conducted using both biographical data and enrollment observations at a South African University by researchers [1]. The basic methodology was to indicate influence of four biographical characteristics This research tried to identify the influence of gender, spoken home language, home province, and race description on student marks over 4 years [1]. Bayesian was adopted to predict the student outcome using posterior probability. The results provided positive sufficient evidence about the effect of biographical data but even though the influence gender, home language, home province, and race description were not thoroughly understood [1].

Some factors of accumulating an F are poor background knowledge in the field of study, very low grades and the incapacity of passing an examination, lack of financial resources. The information needed for a branch of the tree C_i is I(p_i, n_i). Bayesian formalism determines probability that objects has value A_i of A Investigating its allocation of values of A in C as just a feature of their group/class.

7) Conclusion: The contribution of historical information is unclear for students' aggregate performance. It somehow depends on variables considered for students and methods used because under certain conditions it contributes very little and it becomes insufficient. This chapter attempted to give a proper background of our study and other related study areas by introducing our models and datasets that have been used, giving a brief view of the domain, student models that have been put under study by other researchers, investigations then finally concluded by considering logic based model background work.

III. Research Methodology

A. Introduction

In this paper I attempt to use background, individual, and university registration information as features, to predict the following possible outcome: Qualified(Q, "PCD", "MFC", "M2C", "MCC", "M1C", "MBP", "MFA", "MFO","MAP", "M2A", "MCA", "M1A"), Excluded("M2E", "MBR", "MBZ", "RET", "MRNM", "MFE", "MCE", "M1E", "MFL", "DPE", "FTC", "MEXL"). I trained several machine learning predictive models from different archetypes of machine learning to deduce the learner into these two outcome profiles. Confusion matrices are used to model performance and entropy feature selection analysis was performed to rate each features contribution to predicting the class label.

In the previous chapters we have covered the work under this study field and also the background of our research. In this section we will put our focus on the methodology branched out to the design of our research, the construction and its limitations. The methodology consists of three steps: scooping up the applicable features of the problem under review and preparing the data, constructing the classification models and its evaluating, and finally the issues of extrapolation student outcomes.

B. Research Hypothesis

The accuracy of predicting a student's outcome can be predicted beforehand using the various algorithms such as Naive Bayes, Decision Tress, Linear regressions ,neural networks e.t.c. Therefore, can we apply the decision tree algorithm and neural networks using Degree enrollment details and the students marks as features to improve the prediction of the outcome code of a specific student?

1) Research Question(s): In order to address the purpose of this research, the study will seek to answer the following questions:

- Can the machine learning algorithms be modelled from the university data?
- Can machine learning algorithms be used for prediction of a student's academic performance to optimize the prediction of student outcome?

Then we should be able to identify the features artefact that can be designed to automatically predict academic performance of a specific student.

C. Data Collection, Pre-processing, and Ethics

The data used in this study is obtained from the Academic Information Systems Unit (AISU) at the University of the Witwatersrand, Johannesburg. It consists of biographical and enrollment observations of students from the Faculty of Science at a Research Intensive Higher Education Institution. The enrollment and biographical observations is limited to gender, home language, and race description for all students registered anytime between the years 2015 to 2018. Instances are identified using the encrypted student number for confidentiality infringement. My research is concerned with optimizing student prediction to discover significant information originating from educational data.

data mining deals with significant information discovery in data stored in databases or information repositories. Its main goal is to find significant structures or piece of information in data that can improve data usage.

Availability and storing of data electronically has been commonly used in past years as well turning such data into knowledge that is useful also information for huge implementations. The application includes artificial intelligence, machines learning,and other important field of study. Therefore, I will extract important features based on Entropy measures from decision trees ,then we will pre-process this data to remove all the noise.

This process can be applied in any size of database because searching for meaningful information requires delicate applications of these algorithms ID3articleOgunde. Nonetheless, due to this enormous growth of data, the need and requirement for sufficient system for these computations is required.

D. Database evaluation

The dataset consists of the following features (CalendarInstanceYear,ProgramFacultyName,AcademicCareer,ProgramCode, ProgramTitle,PlanCode,PlanDescription,YearofStudy, RegistrationStatus,ProgressOutcomeType, ProgressOutcomeTypeDescription,NewtoProgram, NewtoUniversity,CourseCode, CourseTitle,FinalMark ,FinalGrade,EncryptedStudentNo,Gender,RaceDescription ,Age,MaritalDescription,HomeLanguageCode, HomeLanguageDescription, ReligionDescription,CitizenshipStatus, NationalityShortName,CountryOfBirth, PermitTypeDescription,SimsUrbanRural,SimsQuintile, SecondarySchoolQuintile, SecondarySchoolCode,SecondarySchoolName,HRAddress1 ,HRAddress2,HRAddress3, HRAddress4,HRCity,HRPostalCode, HRProvince,HRCountry,HRCountryName,HPAddress1, HPAddress2,HPAddress3,HPAddress4,HPCity,HPPostalCode, HPProvince,HPCountry,YOS1MajorOne, YOS1MajorTwo,YOS1MajorThree,ProgressoutcomeYOS1 ,YOS2MajorOne,YOS2MajorTwo, YOS2MajorThree,YOS3MajorOne,YOS3MajorTwo ,YOS3MajorThree,AggregateYOS1,AggregateYOS2, AggregateYOS3), these feature were then reduced firstly based on elimination of some features such as gender ,race e.t.c(ethnicity features), then secondly, using entropy evaluation feature reduction to remove all unnecessary attributes that do not contribute much to our analysis. All numerical attributes were normalised according to Gaussian distribution normalization taking into account of the categorical variables. All quantifiable features are in between 0 and 1. Missing variables were not catered for, so instead i removed them to avoid excessive noise.

E. Conclusion

I have highlighted the method used to gather the dataset, and explained the processed used to finalize and process the dataset. Therefore, about 30% of the entire data was successfully processed and 10 folds was applied on each model used on the processed dataset.

IV. Modeling and consistency validations

A. Introduction

Above, we have explained the steps taken to process our dataset. In the following steps, we will span our problem space for attributes evaluation prediction, identify the transactions between possible pass and fail ground truth of each evaluation code, and then finally the evaluation of our models. This paper presents multiple algorithms for scaling up the classification accuracy of Multi-Neural Network wiht logistic regression, decision tree and naive Bayes classifiers in multi-class classification problems. Amongst other data mining methods, the selected algorithms have some advantages: (a) transparent, (b) implementable without extra effort, (c) less or no prior knowledge required, (d) ability to handle both numerical and categorical data, (e) vigorous, and (f) can withstand noisy data. A naïve Bayes classifier is a simple probabilistic classifier based on: (a) Bayes theorem, (b) strong (naïve) independence assumptions, and (c) independent feature models. It is also an important mining classifier for data mining and applied in many real world classification problems because of its high classification performance. They all have several advantages such as (a) easy to use, (b) only one scan of the training data required, (c) handling missing attribute values, and (d) continuous data.

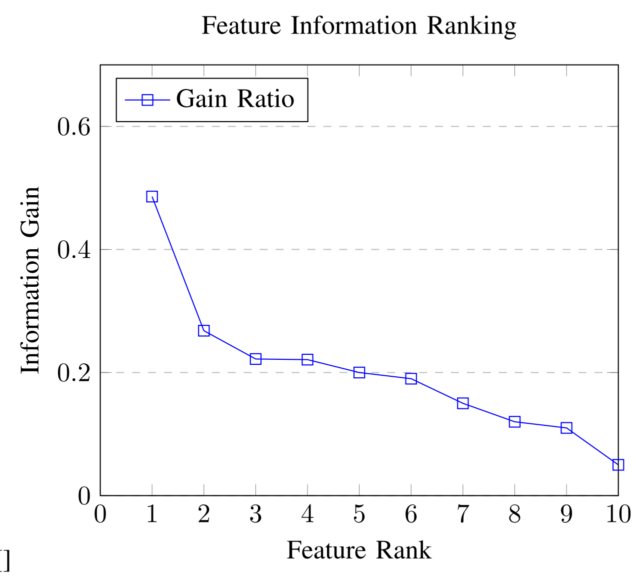
B. Problem space for Student Prediction analysis

The result of entropy evaluation using both the information gain and gain ratio resulted in eliminating some of the attributes used such as Year of study three , this reasoning is based on the fact that I am trying to predict the possibility of a student failing or passing which is the outcome of a student in the preceding year, therefore since the Third Year is the final year we didn't need it to predict if the student will pass the upcoming year. Another one was the 3rd major from second year(which is Computational and Applied Mathematics(CAM)), this elimination is based on the fact that we have less student registered for CAM in 3rd year because it is actually more challenging than Maths and Coms. The below curve(in the following section) show the entropy evaluation of the features under investigation, I reduced the visualization eliminating all features that had 0 contribution leaving only the top 30 features (namely :CalendarInstanceYear, ProgramFacultyName, AcademicCareer,ProgramCode, ProgramTitle,PlanCode,PlanDescription, YearofStudy,RegistrationStatus,ProgressOutcomeType, ProgressOutcomeTypeDescription,NewtoProgram, NewtoUniversity,CourseCode,CourseTitle, FinalMark, FinalGrade,YOS1MajorOne, YOS1MajorTwo,YOS1MajorThree,ProgressoutcomeYOS1, YOS2MajorOne, YOS2MajorTwo, YOS2MajorThree,YOS3MajorOne,YOS3MajorTwo, YOS3MajorThree,AggregateYOS1,AggregateYOS2,AggregateYOS3) but I will only visualize the 1st 10.

1) Attributes acquisition and configuration: The attribute selection method is a data mining procedure which estimates the utility of attributes for given task (usually prediction). Attribute evaluation is used in many data mining tasks, for example in feature subset selection, feature weighting, feature ranking, feature construction, decision and regression tree building, data discretization, visualization, and comprehension.

I used two methods namely (*Feature Gain Ratio*) feature information ranking and (*Feature Information Ranking*) entropy evaluation

After the dataset was preproccessed under-going the

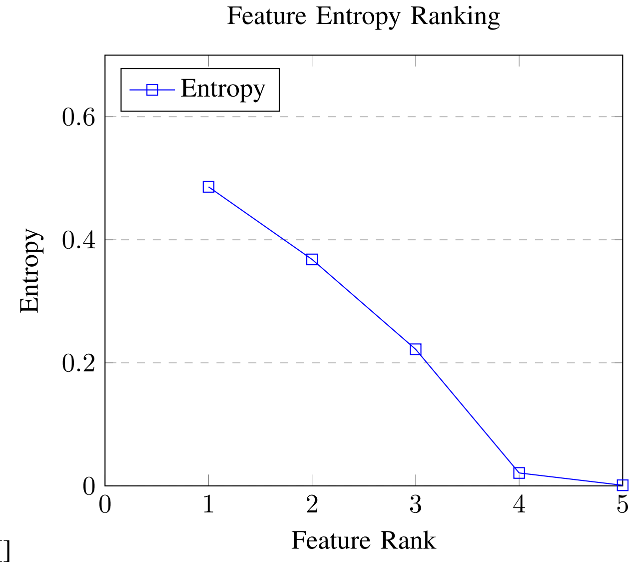


[]

cleaning process and information gain evaluation for attribute selection, I remained with 23 attributes, resulting in narrowing our focus and eliminating features that didn't contribute much or at all from unnecessarily noising the data. The identification of the contributory parameters was done by performing sensitivity analysis that was conducted over the entire training set and measured the effects of small changes in each of the input parameters as they relate to the output. This made it easier for our model to predict outcome based on important features only.

Therefore, each new entry can now contain only 23 features instead of the original 52/53 + 9(custom-made) featured attributes in the original dataset.

From these remaining 23 features I applied entropy



[]

evaluation, and this process eliminated more features further , remaining with only 5 electable ones (YOS1MajorOne, YOS1MajorTwo,YOS2MajorOne, ProgressoutcomeYOS1,ProgressoutcomeYOS2) in the figure above. The most important feature are the 1st two 1st year majors(mathematics and Computational and Applied Mathematics) ,followed by the Computer Science from 2nd year and the 1st and 2nd year progress outcomes. This makes sense since our main point is to predict the outcome of the preceding year.

C. *Evaluation of the Model*

Before I start training any of the machine learning's classification algorithm first I had to look at some constraints and performance metrics to follow :

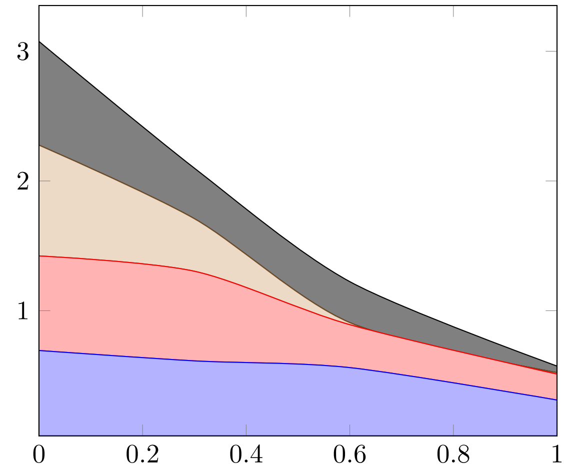
1) *Constraints:*

- low-latency requirement.
- Interpretability is not that much important.
- Errors cannot be very costly.
- Probability of a data-point belonging to each class is needed.

2) *Performance Metric(s):*

- Multi class log-loss
- Confusion and precision matrix

and the Loss Curves for all 4 classification options is as follows:



D. *Confusion Matrix Of Each Model*

While most model used in this report performed really good, Bayesian networks performed at second lowest with overall accuracy of 71, this is a type of probabilistic graphical model that uses Bayesian reasoning based on weight. Its point of purpose is to model based on subject to one or more conditions or requirements being met, dependence, and therefore representing conditional dependence using graph edges.

Basically, the range of the log loss is [0, *inf*] and the goal of our machine learning models is to minimize this value. A perfect model would have a log loss of 0. The value of log loss increases as the predicted probability diverges from the actual label, but for a certain value of log loss other than 0, can we quantify how well is our model is performing. The data was decretized to numeric so that it can allow my Feed forward Artificial Neural Network to be able to train with all 23 features before entropy reduction and information gain evaluation.
My Feed forward Artificial Neural Network with regression model showed a positive performance where an error was minimised from

- epoch = 10, MSE = 0.614, TRAIN ACC = 0.350
- epoch = 20, MSE = 0.579, TRAIN ACC = 0.613
- epoch = 30, MSE = 0.518, TRAIN ACC = 0.70

.
.
.

- epoch = 470, MSE = 0.520, TRAIN ACC = 0.759
- epoch = 480, MSE = 0.419, TRAIN ACC = 0.764
- epoch = 490, MSE = 0.311, TRAIN ACC = 0.774

From above we can see that our algorithm improved its classification minimizing the error to 0.019 with a training accuracy of 77.4%

My test student instance was predicted correctly with an average of 71% confidence. So far the models are doing good except for the Bayes net since this had to do with out dataset being numerical and categorical, Therefore, I can conclude that it is possible to predict student outcome using previous enrollment figures for a specific year.

For this algorithm to run, I had to eliminating the parameters that are the least contributory to successful entropy steps mentioned above. These steps reduced the input noise improving the network's performance. Converting the categorical attribute features to numerical discrete values and all the input parameters that were not contributing much to the output could be considered superfluous and thus eliminated from the input space.

		Predicted		
		Qualified	Excluded	Proceed
Actual	Qualified	304	46	3
	Excluded	51	132	28
	Proceed	1	1	58

Correctly Classified Instances 501 which is 80.28% and 123 where Incorrectly Classified Instances which is 19.71%

		Predicted		
		Qualified	Excluded	Proceed
Actual	Qualified	438	38	2
	Excluded	70	62	9
	Proceed	1	3	1

The CART algorithm takes into account both Classification and Regression Trees. It is a binary decision tree that takes a node and split it into two child nodes repeatedly constructing child nodes from the root node that has the whole learning sample. Correctly Classified Instances 512 which is 82.05% and 112 where Incorrectly Classified Instances which is 17.94%

		Predicted		
		Qualified	Excluded	Proceed
Actual	Qualified	446	32	0
	Excluded	74	66	1
	Proceed	1	4	0

The hybrid decision tree is able to remove noisy data to avoid over-fitting. The hybrid Bayes classifier identifies a subset of attributes for classification. Both algorithms are evaluated using 10 real benchmark datasets. They outperform traditional classifiers in challenging multi-class applications.

		Predicted		
		Qualified	Excluded	Proceed
Actual	Qualified	446	32	0
	Excluded	74	66	1
	Proceed	1	4	0

E. Reliability of prediction data

TP-Rate	FP-Rate	ROC-Area	Class
0.54	0.27	0.76	2
0.56	$9.78 \cdot 10^{-2}$	0.61	1
0.64	0.36	0.69	0

Some prediction where similar in contexts, therefore, the algorithm did not classify all instance correctly. These results although produced a minimum accuracy of **TEST ACC = 0.54** with a max-prediction accuracy of 0.821.

F. Model Differences

In order for our classification to work efficiently, we model subjects of the same data with its default non-normalized classes, so that we don't get to have a large variance in our datasets class distribution.

V. Conclusion

In this work, I presented a new approach to the multi-label progress outcome classification task. First, In CART and ANN with Regression, I proposed a transformation method to transform the problem into a single binary classification problem. Afterwards, I developed a deep learning-based system to solve the transformed problem. The key component of our system was the embedded models namely Logistic Regression and Feed Forward Artificial Neural Networks, which used all 23 features to predict the outcome. This system overall performed very well with an accuracy of upto 82.3%, achieving a most likely the same score that was achieved by [13] for multi-label emotion classification problem. I found that the student outcome prediction can be improved by further finding the principal Component function that can model the relationships between the input train dataset and the labels without reduction, which helps to improve the system's performance. Moreover, this will show that the system is interpretative by visualizing the principal Component spanned on at-most on a 2-*d* which will be easy for analyzing them . These results showed that my system can perform even better since the dimension of our training set can be reduced further instead of having all 62 features before data cleaning took place. However, some limitations have been identified. Our system does not model the relationships between student of different course, schools and the labels, because our dataset was custom normalized to fit our models with one contrast structure. Thus, in our future work, I plan to work on solving this drawback. One possible solution is to adapt the attention function to model the relationships between different schools, ethnicity,race e.t.c. In this case, SVMs may be useful in identify individual class and processed for validation. [14].

REFERENCES

1 Ajoodha, R. and Jadhav, A., "Identifying at-risk undergraduate students using biographical and enrollment observations for mathematical science degrees at a south african university," *Private Communication*, vol. 1, no. 1, pp. 1–21, nov 2017.

2 Considine, G. and Zappalà, G., "The influence of social and economic disadvantage in the academic performance of school students in australia," *Journal of Sociology*, vol. 38, no. 2, pp. 129–148, 2002. [Online]. Available: https://doi.org/10.1177/144078302128756543

3 Quinlan, R., *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

4 Al-Radaideh, Q., Al-Shawakfa, E., and I. Al-Najjar, M., "Mining student data using decision trees," *The International Arab Journal of Information Technology - IAJIT*, 01 2006.

5 Kovacic, Z. J., "Early prediction of student success : Mining students enrolment data," 2010.

6 Kotsiantis, S. B., "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007, pp. 3–24. [Online]. Available: http://dl.acm.org/citation.cfm?id=1566770.1566773

7 Minaei, B., Kashy, D., Kortemeyer, G., and Punch, W., "Predicting student performance: an application of data mining methods with an educational web-based system," vol. 1, 12 2003, pp. T2A– 13.

8 Quinlan, J. R., "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

9 Dietterich, T. G., "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*, ser. MCS '00. London, UK, UK: Springer-Verlag, 2000, pp. 1–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=648054.743935

10 Sison, R. and Shimura, M., "Student Modeling and Machine Learning," *International Journal of Artificial Intelligence in Education (IJAIED)*, vol. 9, pp. 128–158, 1998. [Online]. Available: https://telearn.archives-ouvertes.fr/hal-00257111

11 Herzog, S., "Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression," *New Directions for Institutional Research*, vol. 2006, no. 131, pp. 17–33, 1 2006. [Online]. Available: https://doi.org/10.1002/ir.185

12 Mason, C., Twomey, J., Wright, D., and Whitman, L., "Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression," *Research in Higher Education*, vol. 59, no. 3, pp. 382–400, May 2018.

13 Perikos, I., Ziakopoulos, E., and Hatzilygeroudis, I., "Recognize emotions from facial expressions using a svm and neural network schema," in *Engineering Applications of Neural Networks*, Iliadis, L. and Jayne, C., Eds. Cham: Springer International Publishing, 2015, pp. 265–274.

14 Mirsamadi, S., Barsoum, E., and Zhang, C., "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2227–2231.