Learning Analytics in Further and Higher Education

Jade Lyons[1], Kevin Curran[2]
Northwest College of Further and Higher Education[1]
School of computing and Intelligent Systems, Ulster University[2]
Northern Ireland

Abstract

Student retention is a problem faced by all education institutions. We demonstrate how we applied machine learning techniques to data about students, their courses, attendance records and whether they dropped out. We highlight how we used data analysis and visualization techniques as predictors for student drop out. We found that Random forest shows good results but specificity is low. More data would probably resolve this issue. Although neural networks has next greatest accuracy, k-NN's results over the different descriptors are reasonable. Specificity is quite low over all models. More work would need to be carried out to achieve higher accuracy. As the data was so imbalanced, it would be better to have a bigger dataset so oversampling would not be needed. This would help to achieve better results.

1. Introduction

As we have been living in the digital age for years now, large amounts of digital data has been accrued by schools and universities about students and the potential of this data is now beginning to be realized. Various government and education initiatives exist that are encouraging schools and colleges to make effective use of the data that is being held about their students, to enhance the student experience, identify 'at risk' students before they might leave and therefore improve retention. Applying data analysis techniques can help schools create strategies that will help them engage those students who may otherwise have dropped out, as well as making the education experience better for all students [1,2].

There are a great number of research papers available that have looked at how retention can be improved and if there are any specific predictors that make it more likely that students will drop out [3,4]. There are a few similar characteristics that seem to run through a lot of this research. John Bean, who has studied retention extensively since the 80's and has developed a psychological model of retention, suggests that factors in a student's background can affect whether they are likely to drop out or not [5]. Environmental factors and what the student's intentions are, also play an important role as indicators of a student's commitment to a course of study. Assuming that this is an accurate theory, then each individual college will be looking at their specific factors that have an impact on student retention. Crucial to guiding the development and implementation of measures to improve student retention at an institution is an understanding of the factors that influence retention at that institution [6,7].

The term Learning Analytics has been applied to the analysis of data held about students. Learning Analytics is "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [8].

They have published several documents about learning analytics and provided online learning resources to help schools who want to start making use of their student data to enhance their own teaching and learning [4]. In a wider context other than improving retention, the National Forum have stated that Learning Analytics can be used to enhance teaching and learning in a variety of ways:

- It can let teachers know which resources their students are using and how active they are.
- It can let students know how engaged they are with course material, relative to their peers.
- Real-time information can give both teachers and students the opportunity to take timely, informed action, as appropriate.
- It can inform the curriculum and programme design.
- It can identify patterns of activity that are most likely to engender deep learning and have a successful outcome for the student.
- It can identify at-risk students and empower them to change their academic trajectory before they suffer any negative consequences.
- It can identify and prescribe actions and resources that are most likely to yield a favourable outcome for students.
- It can be used to identify students with sudden changes in engagement that can be indicative of a wide range of non-academic issues. By identifying students that may be facing personal, emotional, medical, social or financial challenges, LA can help support staff to proactively intervene and provide relevant, targeted supports to students with the greatest need [9].

Learning Analytics can be divided into three levels: descriptive, predictive and prescriptive [10, 11, 12].

Purdue University used a predictive model and developed an early alert system that is an algorithm that processes the data about each student and uses a traffic light system to alert the teacher of 'at risk' students as well as alerting the student that they may experience difficulties completing that module. We look at the student data that is held by the North West Regional College over a three-year period.

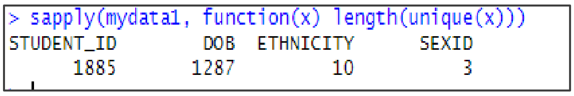
2. The Data

The North West Regional College has provided seven csv files with a variety of data relating to students and courses. This data has been harvested from a few different MIS systems within the college. The files, attributes and number of observations are as follows:

- Student.csv: STUDENT_ID, DOB, ETHNICITY, SEXID, LAST_NAME, TUTOR_STAFF_ID, MILESTOCOLLEGE (1890 observations)
- Course.csv: COURSE_ID, SUBJECT, TITLE, COURSE_AIM, INST_TIER1, INST_TIER2, INST_TIER3, TENANT_ID, AWARDING_BODY (68 observations)
- Course_instance.csv: COURSE_INSTANCE_ID, COURSE_ID, START_DATE, END_DATE, ACADEMIC_YEAR (170 observations)
- Staff.csv: STAFF_ID, FIRST_NAME, LAST_NAME, TITLE (182 observations)
- Staff_course_instance.csv: STAFF_ON_COURSE_INSTANCE_ID, STAFF_ID, COURSE_INSTANCE_ID (1039 observations)
- Student_course_membership.csv: STUDENT_COURSE_MEMBERSHIP_ID, STUDENT_ID, COURSE_INSTANCE_ID, WITHDRAWAL_REASON, COURSE_OUTCOME, COURSE_EXPECTED_END_DATE, COURSE_JOIN_DATE, COURSE_JOIN_DATE, COURSE_END_DATE (1890 observations)
- Attendance.csv: EVENT_ID, STUDENT_ID, EVENT_NAME, START_TIME, END_TIME, EVENT_ATTENDED, COURSE_INSTANCE_ID (799995 observations)

At first glance there appears to be relationships between each of the csv files as there appear to be primary and foreign key fields throughout the files. As there is a lot of information here, each file was looked at in turn to identify relationships and to decide which files will be used in the analysis. For the purposes of this analysis the following files were selected as the most use for determining student drop out: student, attendance and student_course_membership. These files can be joined on STUDENT_ID.

The student data frame was examined first, to check how many unique instances of each attribute were present. Interestingly, although the student data frame showed that it had 1890 observations which would suggest that there were 1890 individual students' records, there were only 1885 unique observations:



```
> sapply(mydata1, function(x) length(unique(x)))
STUDENT_ID        DOB  ETHNICITY      SEXID
      1885       1287         10          3
```

Figure 1. Unique observations in student data

There were five duplicates that were removed. Even if a student has completed more than one course they should only have one student record in this file. All the attributes of the student data frame should be useful in the analysis. STUDENT_ID will be used as the key to join the three selected files together. DOB is useful at this stage as it can provide a student's age. SEXID is useful as knowing whether more males or females drop out is interesting. MILESTOCOLLEGE will be a useful predictor as there may be a relationship between the distances that people have to travel and whether they are more likely to drop out or not. ETHNICITY may be useful if only to show if there is much diversity at the college. Figure 2 shows the number of each ethnicity of the students in the student file.

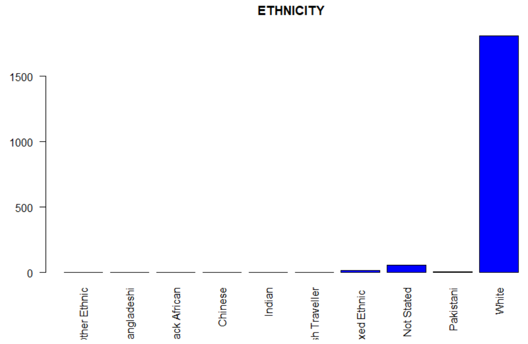


Figure 2. No of each ethnicity

From this plot it is obvious that students are mostly identifying as white. As there is such a clear difference in the number of white students compared to other ethnicities, this will probably not make a good predictor for the model.

The attendance file has a number of very useful attributes for this analysis. Every student's attendance at every class within a course has been recorded as either 'Y' or 'N'. Using this data, two new columns of data were created, ABSENCES and CLASSESATTENDED. These represent the total number of absences and classes attended for each student. These new attributes were then added to the student file, joined on STUDENT_ID. The data frame now has 3503 observations. Student_course_membership was examined in the same way and five duplicate student observations were removed which resulted in 1886 observations. Student and student_course_membership were linked via the STUDENT_ID, this resulted in 3503 observations and 16 attributes in the new data frame. There were 8 attributes that were removed from the data frame as they weren't useful as predictors. The COURSE_OUTCOME attribute reflected the WITHDRAWAL_REASON attribute with a different code that is used internally for management. This attribute was not needed for this analysis and was removed. There were other attributes that weren't useful for this analysis as they had no value as predictors, they were: COURSE_INSTANCE_ID, COURSE_JOINE_DATE (all the same each year), DOB (age observation exists), COURSE_EXPECTED_END_DATE, COURSE_END_DATE (these are the same for all courses) and STAFF_ON_COURSE_INSTANCE_ID as this is made up of COURSE and STAFF_ID.

The other files were discarded at this time for a number of reasons. They held duplicate data and a lot of the STUDENT_ID's didn't match the IDs in the student data frame. The COURSE_INSTANCE_ID which occurs in a few of the files did not have the same data format in each file so was of no use when trying to join the different files.

There is now one data frame made up of student, student course membership and attendance. The STUDENT_COURSE_MEMBERSHIP_ID attribute's data is made up of the student id and a course id:

FDFS3WELFR_ABB17107726

The course id (before underscore above) is useful as a predictor as students may drop out of certain types of courses more than others. The student id part of the data is no use and was removed, leaving each row of data as only the course id. This new attribute was called 'Course'. Figure 3 shows the attributes of the new merged data frame and the number of missing values (3503 observations):

Attribute	NA no
ETHNICITY	1618
SEXID	1618
Course	1618
WITHDRAWAL_REASON	1618
MILESTOCOLLEGE	1618
COURSE_JOIN_AGE	1618
CLASSESATTENDED	1038
ABSENCES	1038

Figure 3. Missing data in data set

There are a total of 3503 observations in the data frame and they account for nearly half of the data. Because there are so many, it was decided it wouldn't be appropriate to impute values because it could result in inaccurate results when building the models. For this reason the observations with missing values were removed leaving 847 complete observations which, although a lot smaller, will have a lot more integrity when building the models.

Figure 4 is a table plot that gives an overall view of the data that will be used in the analysis.

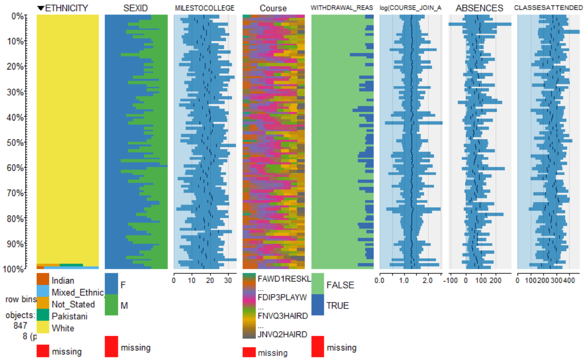


Figure 4. Table plot of dataset

From this plot it is obvious that the most popular ethnicity in this data frame is white. The plot also shows that there is an even spread of male and female students. It also doesn't look like anyone lives extremely far away from the college either. The withdrawal reason column shows that in this data frame, most of the students stayed until the end of the course. The WITHDRAWAL_REASON attribute's data may contain one of 37 codes that represent the reason why students leave, ranging from C01, 'college terminated because of attendance' to U01,'unknown reason'. The codes are only present if a student has left the college. Figure 5 shows which withdrawal code was used the most often for students leaving.

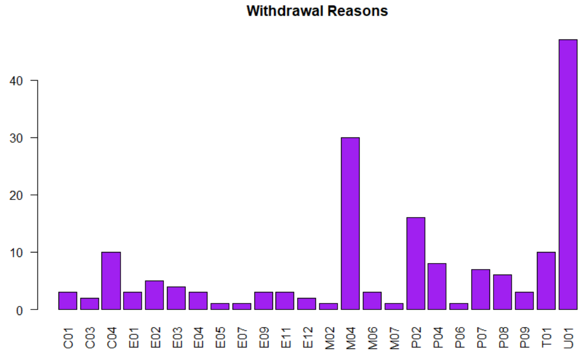


Figure 5. Total number of each withdrawal reason

The most common reason for students leaving was U01 which is 'unknown', followed by M04, 'gone into employment', PO2, 'family/personal issues' and C04, 'college terminated attendance unable to make contact'.

There are other plots that can be looked at to observe any relationships that may exist within the data. Figure 6 shows age and no of absences of the students.

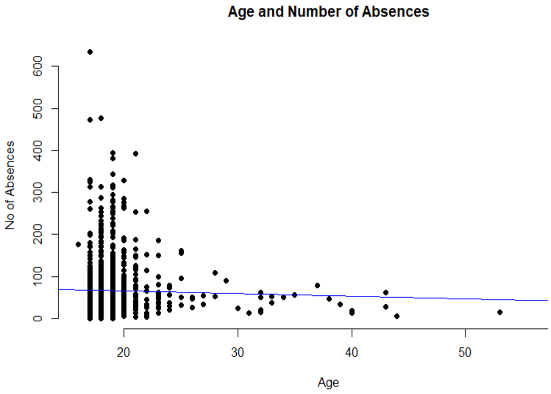


Figure 6. Plotted age and number of absences

The blue line is the regression line which shows how much of a relationship there is between these two attributes. In this case age has a little impact on the number of absences a student can have. Figure 7 shows the miles a student has to travel to college and the number of absences.

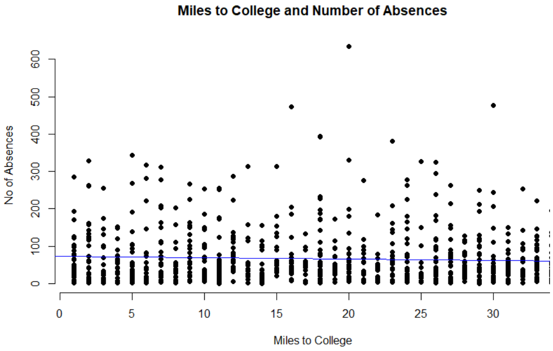


Figure 7. Plotted Miles to College and number of absences

The relationship between these two attributes is even less than the previous one. The number of absences is clearly not very dependent on how far a student lives from the college. Figure 8 shows the relationship between a student's age and how many classes they attended.

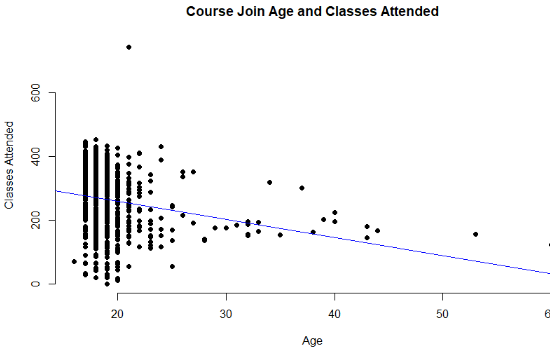
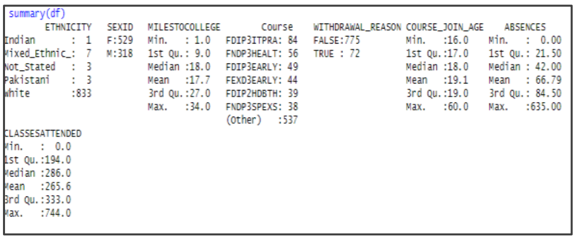


Figure 8. Plotted course join age and number of classes attended

This plot shows that there is some correlation between the number of classes attended and a student's age. The regression line is sloped which shows that a change in the x axis value (age) can have an effect on the y axis (classes attended). The data is skewed, as the majority of students in the data set are younger than 25.

Most machine learning algorithms use numerical data. The new data frame at present is made up of 4 character attributes and 4 integer attributes. The character attributes need some work to enable their use in the analysis to follow. The target variable WITHDRAWAL_REASON is changed to a factor type, a student has either stayed or left the course early, giving two possible outcomes, a factor type is most suitable for this type of data. The other 3 character variables are, ETHNICITY, SEXID and COURSE. These are categorical data because they are not continuous and have a finite number of values. Categorical data falls into two categories, nominal and ordinal. As there are no relationships between the values within the 3 variables, they are considered nominal. Dummy encoding was used to convert the variables, as this was the most appropriate method to deal with categorical variables that are going to be used in analytical methods that require numbers. Dummy encoding creates a new attribute for each of the 'categories' within the data. As a result of the dummy encoding, there were now 74 attributes.

Figure 9 shows that there is a wide range of numerical values in the data. These very different ranges could potentially skew the results of any analysis carried out. The built-in scale function of the CARET package will be used when applying the machine learning algorithms.


```
summary(df)
       ETHNICITY    SEXID   MILESTOCOLLEGE          Course     WITHDRAWAL_REASON COURSE_JOIN_AGE      ABSENCES
Indian        :  1   F:529   Min.    : 1.0   FDIP3ITPRA: 84   FALSE:775         Min.    :16.0    Min.   :  0.00
Mixed_Ethnic_:  7   M:318   1st Qu.: 9.0   FNDP3HEALT: 56   TRUE : 72         1st Qu.:17.0    1st Qu.: 21.50
Not_Stated    :  3           Median :18.0   FDIP3EARLY: 49                     Median :18.0    Median : 42.00
Pakistani     :  3           Mean   :17.7   FEXD3EARLY: 44                     Mean   :19.1    Mean   : 66.79
White         :833           3rd Qu.:27.0   FDIP2HDBTH: 39                     3rd Qu.:19.0    3rd Qu.: 84.50
                             Max.   :34.0   FNDP3SPEXS: 38                     Max.   :60.0    Max.   :635.00
                                            (Other)   :537

CLASSESATTENDED
Min.   :  0.0
1st Qu.:194.0
Median :286.0
Mean   :265.6
3rd Qu.:333.0
Max.   :744.0
```

Figure 9. Summary of dataset

3. Building the models

When the dataset was prepared there were 847 observations and 74 variables (attributes). As WITHDRAWAL_REASON is our target variable, there are 73 variables that can be used as predictors. The algorithms that are being used are:

1. K Nearest Neighbour
2. Logistic Regression
3. Random Forest
4. Neural Network
5. SVM

A confusion matrix will be used to measure the performance of each of the models. They are composed of true positives, true negatives, false positives and false negatives. From the confusion matrix it is possible to measure a models accuracy, sensitivity, specificity, positive predictive value and negative predictive value. The results can then be compared to decide which model is the most accurate in predicting if a student will drop out or not.

N=x	TRUE	FALSE
TRUE	TP	FN
FALSE	FP	TN

Figure 10. Confusion matrix

A. *K Nearest Neighbour*

k-NN is used for making classifications and predictions. As k-NN calculates the distance between datapoints the attributes in the data frame need to be numerical. ETHNICITY, SEXID and STUDENT_COURSE_MEMBERSHIP_ID were all chr attributes. These attributes were converted to ints. Any integer attributes in the data frame were scaled to make sure that k-NN can work accurately.

The data is divided into two. 67% is used to train the selected model and the other 33% is used to test the dataset to verify the validity of the selected trained model. It is important to check if there are any near zero variance predictors. If there are then the model may crash or become unstable. As a result of this check, the variables were reduced to 12. There are different opinions whether removing near zero variances is actually removing valuable data. The technique used here was 'nearzerovar', part of the CARET package. It may have been better to use a different technique. After running k-NN using the CARET package, figure 11 is the confusion matrix that was generated.

N=278	TRUE	FALSE
TRUE	253	23
FALSE	2	0

Figure 11. Confusion Matrix k-NN

The confusion matrix shows that the classifier predicted 'Yes' 255 times and 'No' 23 times. True positives are when the predicted value and the actual value are true. In this result there were 253 true positives out of a total of 278. True negatives are when the predicted value and the actual value are false, when false has been correctly predicted. In this result 0 predicted values were correctly predicted as false. False positives (Type 1 Error) are when the predicted value chosen is true but it is in fact false. There were 2 false positives. False negatives (Type 2 Error) are when the predicted value chosen is false but it is in fact true. There were 23 false negative prediction here. The accuracy can be calculated by adding the true positives and true negatives and dividing by the total number of samples used in the analysis. 253/278 = 0.91. This model was 91% accurate. Sensitivity (recall) is measured by calculating the number of correct positive predictions divided by the total number of positives plus false negatives. The best is 1.0, the worst is 0.0. This model: 253/276 = 0.92. Error rate is calculated by dividing all of the incorrect predictions with the total number in the dataset. In this case, 25/278 = 0.08. (Best error rate is 0.0 worst is 1.0) Specificity (True Negative Rate) is calculated as the number of correct negative predictions divided by the total number of negatives. The best is 1.0, the worst is 0.0. This model: 0/23 = 0. Precision measures how precise a model is in predicting true outcomes. This is true positives divided by total number of predicted positives. The best is 1.0 and worst is 0. This model: 253/255=0.99. The best value for k chosen by the model was 7.

As there were no true negatives and specificity was 0 there could be an imbalance in the training data. An imbalance is more likely to occur in binary classification problems where there is many more of one class than another. By using the 'table' function on the target class, the following result was given:

```
FALSE  TRUE
 520     49
```

This clearly shows an imbalance in the data that can lead machine learning algorithms to struggle with accuracy. To counteract this the ROSE package was used. As this is such a small dataset, oversampling was chosen. After using the table function again the results were:

```
FALSE   TRUE
 520     520
```

k-NN was run again with the new oversampled data and the figure 12 shows the resulting confusion matrix.

N=278	TRUE	FALSE
TRUE	184	12
FALSE	7	11

Figure 12. Confusion Matrix k-NN

The value for k that was chosen was 5.

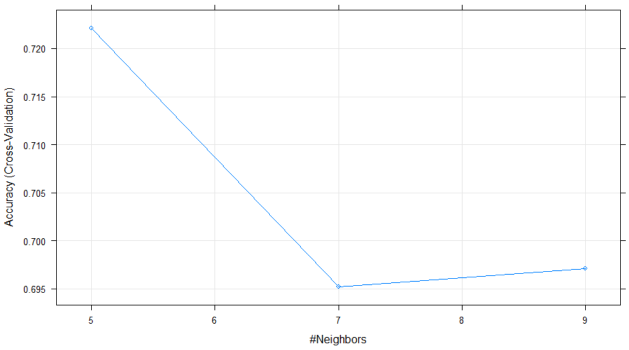


Figure 13. k-NN plotted

Accuracy: 0.70
Sensitivity: 0.94
Error Rate: 0.07
Specificity: 0.47
Precision: 0.96

B. *Logistic Regression*

Logistic Regression is one of the more common machine learning algorithms used for binary classification problems. Figure 14 shows the confusion matrix which resulted from using the

Logistic Regression Algorithm.

N=278	TRUE	FALSE
TRUE	173	9
FALSE	82	14

Figure 14. Confusion Matrix Logistic Regression

Accuracy: 0.67
Sensitivity: 0.95
Error Rate: 0.32
Specificity: 0.60
Precision: 0.67

C. Random Forest

The random forest algorithm is similar to the classification tree aside from the fact that is randomly selects observations and variables from a dataset to build multiple decision trees. Figure 15 shows the results of running the random forest algorithm:

N=278	TRUE	FALSE
TRUE	240	21
FALSE	15	2

Figure 15. Confusion Matrix Random Forest

Accuracy: 0.87
Sensitivity: 0.94
Error Rate: 0.12
Specificity: 0.086
Precision: 0.94

The final value used for the model was Mtry=6. Figure 15 shows variable performance as identified by the random forest algorithm. It shows that the number of classes attended, followed closely by number of absences were the most important attributes to determine if a student will leave a course. Students who attended early year's courses were least likely to drop out.

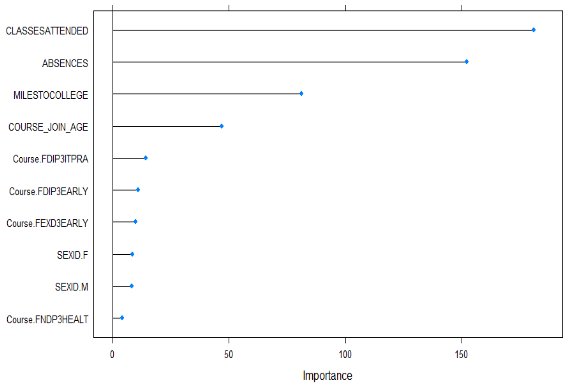


Figure 15. Variable Importance Plot

D. Neural Networks

Neural Networks are a set of algorithms that are designed to recognise patterns in data. Figure 16 is the confusion matrix as a result of using Neural Networks.

N=278	TRUE	FALSE
TRUE	201	16
FALSE	54	7

Figure 16. Confusion Matrix Neural Networks

Accuracy: 0.74
Sensitivity: 0.92
Error Rate: 0.25
Specificity: 0.30
Precision: 0.79

E. Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression purposes when completing data analysis. Figure 17 is the resulting confusion matrix showing the results of using the SVM algorithm.

N=278	TRUE	FALSE
TRUE	174	8
FALSE	81	15

Figure 17. Confusion Matrix SVM

Accuracy: 0.68
Sensitivity: 0.96
Error Rate: 0.32
Specificity: 0.65
Precision: 0.68

4. Conclusion

The student drop out rate anlysis has made use of supervised learning algorithms. This was a classification problem as there were only two possible outcomes, either a student did or did not drop out of a course. Five different machine learning algorithms were applied to the data with the results compared in figure 18.

	kNN	Log Reg	Random Forest	Neural Network	SVM
Accuracy	70	67	87	74	68
Sensitivity	.94	.95	.94	.92	.96
Error Rate	.07	.32	.12	.25	.32
Specificity	.47	.60	0.086	.30	.65
Precision	.96	.67	.94	.79	.68

Figure 18. Results Table

Random forest shows good results but specificity is low. More data would probably resolve this issue. Although neural networks has next greatest accuracy, k-NN's results over the different descriptors are reasonable. Specificity is quite low over all models. More work would need to be carried out to achieve higher accuracy. As the data was so imbalanced, it would be better to have a bigger dataset so oversampling would not be needed. This would help to achieve better results. Ultimately, we show how data analysis and visualization techniques can be used as predictors for student drop out

5. References

[1] Smyth, S., Curran, K., McKelvey, N., (2019). The Role of Education and Awareness in Tackling Insider Threats. Cybersecurity Education for Awareness and Compliance. IGI publishers, pp: 33-52, DOI: 10.4018/978-1-5225-7847-5.ch003.

[2] Duffy, W., Curran, K., Kelly, D., Lunney, T., (2019). An investigation into smartphone based weakly supervised activity recognition systems. Pervasive and Mobile Computing, Vol. 56, pp: 45-56, Elsevier, ISSN: 1574-1192, DOI:10.1016/j.pmcj.2019.03.005.

[3] Al-Masri, A., Curran, K., (2019). Smart Technologies and Innovation for a Sustainable Future, Proceedings of the 1st American University in the Emirates International Research Conference - Dubai, UAE 2017 978-3-030-01658-6, DOI: https://doi.org/10.1007/978-3-030-01659-3.

[4] Curran, K., McNamee, E., McCaroll, N., Chaurasia, P., McBrearty, S., (2019). The Security Considerations in Clou d Adoption for Legal Firms ICSET 2019 – International Conference on Science, Engineering & Technology, Tel Aviv, Israel, 29-30th March 2019.

[5] Bean, J.P., (1990). Why students leave: Insights from research. In D. Hossler , J. P. Bean , and Associates (Eds.), The strategic management of college enrollments (pp. 170-185). San Francisco: Jossey-Bass.

[6] Murtaugh, P.A., Burns, L.D. and Schuster, J., (1999). Predicting the retention of university students. Research in Higher Education 40, 355–371. https://doi.org/10.1023/A:1018755201899

[7] Sadek Mahmoud, M., Eltoweissy, M., Curran, K., and Elmedany, W., (2019). Introduction to special issue on smart systems facts, challenges and technologies. International Journal of Computing and Digital Systems, 8(2), 85-89. [2]. https://doi.org/10.12785/ijcds/080201

[8] McKelvey, N., Curran, K., (2015). Developing Team Work in IT Education to Foster Student Engagement. Recent Advances in Ambient Intelligence and Context-Aware Computing, DOI: 10.4018/978-1-4666-7284-0, ISBN13: 9781466672840,IGI Global Publishing, USA

[9] National Forum for the Enhancement of Teaching and Learning in Higher Education, (2017). Using Learning Analytics to Support the Enhancement of Teaching and Learning in Higher Education," in teachingandlearning.ie, https://www.teachingandlearning.ie/publication/using-learning-analytics-to-support-the-enhancement-of-teaching-and-learning-in-higher-education / (Access Date: 23 February, 2021)

[10] Learning and Academic Analytics, Siemens, G., (2011). http://www.learninganalytics.net/?p=131, (Access Date: 23 February, 2021).

[11] Mc Carroll, N., Curran, K., (2012). Social Networking in Education. International Journal of Innovation in the Digital Economy (IJIDE), Vol. 3, No. 2, pp: 1-15, DOI: 10.4018/jide.2013010101, ISSN: 1947-8305, IGI Global.

[12] Curran, J., Curran, K., (2018). Biometric Authentication Techniques in Online Learning Environments. Biometric Authentication in Online Learning Environments, IGI Publishing. pp: 266-278, ISBN: 9781522577249, DOI: 10.4018/978-1-5225-7724-9.ch011.