Online Student Grade Prediction

Simon Horton
Department of Computer and
Information Science
Fordham University
NYC, USA
sh43@fordham.edu

***Abstract*—With the rise of online learning over the past decade and the necessity of online learning during Covid it will be increasingly important for learning institutions to ensure that they are offering the optimal learning environment for the students. The main objective in this research is to understand the impact that interactions with online materials have on a student's ability to pass or fail a given course. The prediction for the performance of the students used a plethora of models including KNN, Decision trees, Random Forest and majority voting. In the performance analysis it is shown that the random forest is the best performing model for the data. The initial attempt for the 3 combined datasets is to predict pass, fail, withdrawal and distinction. This problem introduces an imbalance to the data that is addressed using SMOTE and under sampling with SMOTE giving the best performance. The results were lower than expected so pass / fail was also investigated, which yielded a better result. The initial aim was to maximize the accuracy and then look deeper to maximize precision, recall and F1 with the predictions. This was maximized using the incremental assessment. When only looking at demographic and VLE interactions, a reasonably accurate prediction was attained for pass fail with all 4 category predictions having a relatively low accuracy and other measures.**

Keywords— Online Learning, VLE, Interactions, classification, Machine Learning, Classifier Performance

I. Introduction

With the online learning market growing by 900% from its inception in 2000 [1] it is becoming an integral part of the future of learning and how the next generation will be educated. There is significant investment also being made in the education space with VC funding growing 6X between 2017 and 2021 [2].

With this growth and investment in the market it will be vital to identify what attracts students and what ensures they stick with the program. There are many factors that students will consider when choosing their online institution such as the quality of the online material, retention, graduation, cost, and curriculum quality [3].

This research will look at the areas that the students interact with and if this is a contributing factor to their success. This will involve using machine learning tools to predict if a student passes or fails the course and looking specifically if this is impacted by the total interactions, the interactions of certain areas and then introducing grades. The reason that grades are introduced is they are seen as a strong indicator of future performance which will be discussed later on in related work.

From the authors perspective having gone through online learning from 2009 – 2022 and having seen the industry change first hand, having a strong platform, access to information and price were all strong requirements for selecting and completing a course.

From an institutional perspective it will be good to understand which areas of interaction yield the most positive correlation to passing and if there are different areas that failing students spend more time interacting with. As such being able to redesign platforms to funnel students to the more beneficial areas and obfuscate or remove the areas that yield more fails.

There is also the cost element and the value for money in aspect. With roughly 62% of post-secondary institutions being nonprofit as of 2021 [4] it is essential that the resources of the institutions are allocated in the most efficient way possible. Analysis such as this will help these institutions to find out which areas are adding the most value to the student success.

The approaches used in this paper will include KNN, Naïve Bayes, Decision Trees, Random forests and an ensemble of different models to try and see if there is away to accurately predict performance based on interactions. Where there are 4 categories to be classified imbalanced data techniques are employed to make it a fairer assessment. There will also be feature selection and validation processes to ensure that the experiment is as thorough as possible. This will be explained in detail in the body of the paper.

The paper will be broken down as follows: II will discuss related works, III will give a detailed view of the data, IV will look at the analysis and findings for the data V will analyse the methodology used for creating machine learning models and then application of them VI will discuss the results and conclusions with VII being an abstract of future work that could be carried out as part of this project.

II. Related Work

Student performance is a widely researched subject that has been analysed for many years. If one were to search the phrase "predicting student performance" on google scholar it would be seen that there are over 5million papers on the subject alone. The field has been researched for well over 100 years and one can see the change in techniques employed and what was looked at from this period to today.

A. Grade prediction early research

The early research is looking at computing correlation between prior testing and performance. For example, in the paper predicting performance in chemistry they look at applying an aptitude test and providing training to predict the score. There is a correlation between the combination of training and aptitude test of 0.59 suggesting it is a predictor for student performance [5]. So here it can be seen that the early research in the field looked at other academic performance to gauge the student performance.

CT-1 = Chemistry Training Examination
CA-1 = Chemistry Aptitude Examination

Name of college	Placement Examination	Number taking test	Correlation Scores vs. grades	Prob. error of "r"
Case School				
Appl. Science	CT-1	166	0.64	0.03
	CA-1	172	.57	.04
	CT-1 + CA-1	156	.66	.03
Univ. Hawaii	CT-1	79	.47	.06
	CA-1	29	.48	.10
Univ. Iowa	CT-1	82	.53	.05
	CA-1	150°	.42	.05
Univ. Saskatchewan	CT-1	19	.52	.11
	CA-1	19	.26	.14
	CT-1 + CA-1	19	.37	.13
Virginia Poly. Inst.	CT-1	281	.70	.02
	CA-1	278	.45	.03
	CT-1 + CA-1	100°	.73	.03

Average correlation Chemistry Training + Chemistry Aptitude = 0.59
Average correlation Chemistry Training = .57
Average correlation Chemistry Aptitude = .44
° Sample

Fig. 1. Picture of correlations with testing, aptitude and score from Cornog & Stoddard [5].

Figure 1 shows the research that the researchers chose to analyse performance of students over a variety of institutions to ensure that they could compare the correlations and not have a single sample. They had reasonable sample size with the exception of Saskatchewan which had lower takers. Given the time period this can be seen as reasonable for the other universities.

As the research progresses over the years there is still a lot of papers looking at the correlation between prior academic performance with a move towards additional factors that could impact on the students' performance. For example, looking at the big 5 personality traits and their correlation with student performance. From this research it is showing that there is some correlation between the traits and student performance [6].

B. Machine learning in student predictions

There are many models that have been introduced in the later research papers and they have been analysed to show the most common set of features and models that have been used. This analysis was done in 2022 [7]. They show that there is a diverse range of models used with an even distribution across the papers studied:

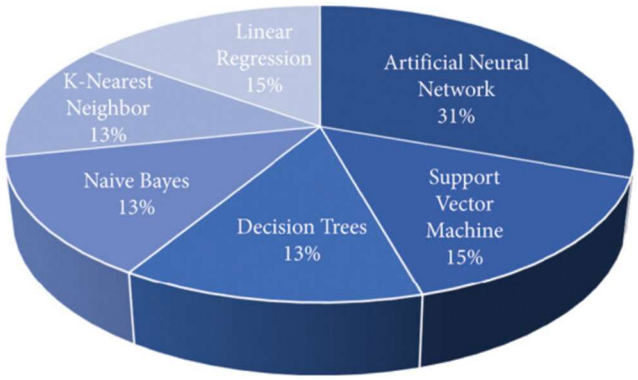

Fig. 2. Distribution of machine learning approaches in student's performance from the analysis [7]

The distribution in figure 2 suggests that the prediction will depend on the type of target that is being predicted. For this dataset it is categorical predictions so the regression will not be a model to consider here.

From the research they show that the top performing metrics for prediciton are Demographic (Student information such as gender, location, family education), Academic (Past test performance, final grades, attendence) and finally Intrnal assessment (interim assessment, study time, plagerism) [7]. These factors ar available in the datasets for this research so it can be shown there is consistency with the historical appraoches.

The research goes on to show the average performance of the different models across several research papers ranging from 80.7% for KNN to 85.9% for an artificial neural network [7]. The type of model used and how the data is prepared is not listed directly in the paper but in the referenced articles so it should be noted that this could have an impact on the performance of the models.

C. Use of student online interactions

With the shown increase in online learning it is important to understand what impact the interactions with online materials have on the outcome for student performance. The 2020 paper Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data [8] shows that alone it is possible to get an accuracy, recall and precision using only these measures. It is apparent that the best measure is still the assessment but combinations of other features will yield comparable results when taken in isolation.

From their research it is shown that assessment, assessment and course access and assessment and mobile course access all produce 99% precision, recall and F1 measure when they use a random forest model [8]. It is shown that course access alone using the random forest produces a score of 97% for the measures where as mobile access is 89% precision, 85% recall and 84% F1. As there are many distractions on the a phone or mobile device it has been shown that roughly 67% of undergraduate students are distracted by their phones [9]. This can come in the form of emails, calls, messages or social media usage. This factor is not present in the dataset for this research paper but important information to consider for further research as the number of distractions increase not only on mobile but on laptops / computers.

D. Research on this dataset

The dataset used for this research was originally used in a 2015 paper titled: OU Analyse: Analysing at-risk students at The Open University [10]. They specifically looked at the at-risk students. They note that demographic data alone can be used when VLE and prior academic performance is not available, but it does not yield the best results. With the introduction of this VLE data it diminishes the value of the demographic data and VLE becomes the predictor of student success [10].

A lot of emphasis in the paper is put on the time of the interactions of the students. For example, they look at forum interactions in the first 3 weeks and gauge if a student will succeed or fail the given courses:

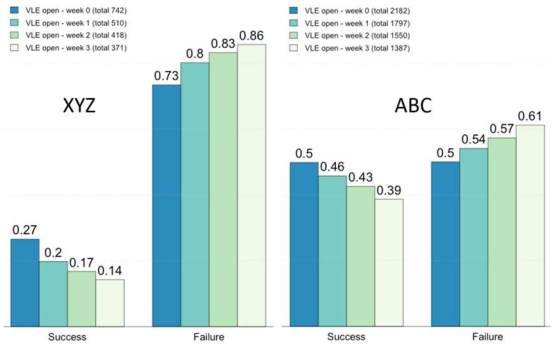

Fig. 3. Analysis of course success or failure when interacting with forums in the first 3 weeks. Page 5 [10].

From figure 3 It is not a clear indication of success or failure in either course as it is showing a high likelihood in XYZ for failure and limited scope for success. The paper goes on to make predictions as more data becomes available in the form of assessment results. It is then more evident that the better the students do on the assessments the better the predictions are. This correlates with the previous research showing that a strong predictor of student success is data of prior assessment performance.

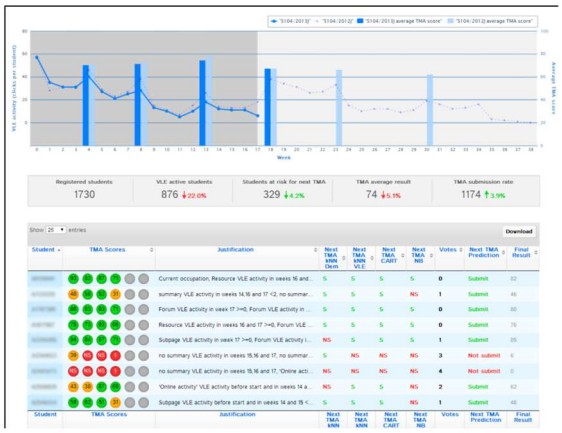

Fig. 4. Extract of dashboard used to monitor student at risk student performance. Page 8 [10]

Figure 4 above shows the monitoring dashboard produced in the research and given to the OU to monitor at risk students for upcoming assessments.

This information combined with the future work to expand the data sets and look to use this monitoring in more courses shows the power of predictive analytics and the way data science can be deployed to enable student success. For this research paper It will look deeper at the prediction side using the available data and applying different techniques / models to see how the performance measures can be optimised as well as identifying key features. It will also look deeper at the demographic factors and those that could be impacting student performance as well as the individual areas that the students are accessing.

III. Data

This section will describe the datasets as well as provide an insight into how the data was combined to produce the analysis and machine learning models to make student predictions. The issue of missing data is also addressed in this section.

A. Description of datasets

The data is available on Kaggle and called Open University Learning Analytics Dataset [11]. It features 7 datasets that include: student registration, assessment, course, student assessment, student information, student VLE and VLE. The table below will list all of the fields across the different data sets. There are 3 common fields that appear across the majority of tables. Where these are featured it will be noted in the header with the addition of "+CP, CM, IDS" for all 3. Alternatively a combination of the 3 if they are featured.

TABLE I. DATASETS AND AVAILABLE DATA

Field Name	*Example*	*Type*
Common Fields		
Code_Module CM	AAA	Text
Code_Presentation CP	2013J	Text
ID_Student IDS	11139	Integer
Student Registration + CP, CM, IDS		
Date_Registration	-150	Integer
Date_unregistered	? - 10	Object
Assessment + CP, CM		
Assessment_type	TMA	Text
Date	20	Integer
Weight	10.0	Float
Course + CP, CM		
module_presentation_ length	299	Integer
Student Assessment + IDS		
ID_Assessment	1752	Integer
Date_Submitted	20	Integer
Is_Banked	0 / 1	Boolean
Score	80	Integer
Student Information + CP, CM, IDS		
Gender	M/F	Categorical
Region	Scotland	Location
highest_education	HE Qualification	Categorical
imd_band	10-20%	Categorical
age_band	35-55	Categorical
num_of_prev attempts	0 / 1	Integer
studied_credits	240	Integer
disability	N	Boolean
final_result	Pass	Categorical
Student VLE + CP, CM, IDS		
ID_site	546652	Integer
Date	20	Integer
Sum_click	10	Integer
VLE + CM, CP		
ID_Site	546652	Integer
activity_type	resource	Categorical
Week_From	2	object
Week_to	2	object

Table one details all of the fields and types that are available for analysis. A brief description of the data will be provided.

Student registration when they registered for the course. Th interpretation here for the date is -150 as in the example means that they registered 150 days before the first day of the course. Date unregistered is for the students that withdrew and on what date. "?" is used as a placeholder for those students that did not withdraw from the course.

For the assessment data it gives details of the assessment types and when they are in terms of the start of the course. For example, 20 would mean that the assessment is 20 days after the start of the course. There are 3 types TMA, CMA, Exam which are Tutor marked assessment, computer marked

assessment and traditional exam respectively. This dataset is combined with interactions and the student assessment to show how the students performed at each stage in the course.

The course dataset shows the length of the course and is not used in the research in its current form. It could be used for deeper analysis of the timings of the courses and assessing the percentage of interactions at different intervals throughout the presentation.

The student assessment dataset is a key set for the research as it shows the students' performance in each of the different assessments. The banked field is for when a student is having another attempt at the course and is using a previous grade for this presentation. The grade is a simple numeric value on the scale of 0 to 100 with 0 being no points and 100 being a perfect score. When combined with the Assessment dataset it enables the necessary data to help in the predictive models. For example, it could be done for the different stages, analysis of the timings of submissions or any other such analysis that can be thought of.

The student information dataset is the key to the research project as it contains the most information about the students as well as the target that the research aims to predict. Most of the fields are straight forward or previously described but the following need a bit more description. Highest education relates to the highest education that the student has undertaken in a UK context. For example, A-Level which is equivalent to advance placement in the US.

IMD band is the index of multiple deprivation from the UK that measures the deprivation of an area [12]. It measures 7 metrics: Income, Employment, Health Deprivation and Disability, education skills and training, crime, barriers to housing and services and living environment. These measures are summarized to provide a banding on a range of 0 – 10 or 0% to 100% where 0 is the most deprived and 10 / 100% are the least deprived.

Age band is a collection of ages, being 0-35, 35-55 and greater than 55 putting the students into bins.

Num of prev. attempts is the number of previous attempts that the student has had for completing the module in question. For the vast majority it is 0 but there are exceptions.

Studied credits is their progress on their degree. The aim is 360 credits for the completion of the degree. So, 0 means they are just starting, 240 means they have completed 2 years and so on. Disability is not described so a yes would represent any disability that the student disclosed.

Finally, the final result has 4 categories Pass, Fail, Distinction and Withdrawn. Pass and distinction represent completing the course successfully with a distinction being higher. A fail is a failure to adequately complete the course and withdrawn means that the student withdrew from the course and did not complete the course. This is used for the target in the machine learning and is reduced to pass fail for some instances to attain more accurate predictions.

The student VLE features all the interactions that the student took while completing the course. This is where the interactions will be derived. The ID site column identifies the area of the VLE that the student has interacted with. The date is the day relative to the start of the course when the interaction took place. The sum click is the number of interactions on that day that the student undertook the interaction.

The VLE dataset details the VLE areas that can be interacted with in the courses. The activity type is aligned with the id site field and that is the area of the VLE that has been interacted with. The week from and to are for specific areas such as quizzes that are only available for a brief period of time. The majority of these have the placeholder value "?" as the resources are available are generally available throughout the whole course.

B. Combinations for analysis and machine learning

For the research a combination of datasets are used for the predictions and analysis of the data. They increase in complexity and granularity throughout the research giving better predictions and understanding of the path to success of the students.

The first dataset is a simple combination of the student information dataset and the summation of the sum click for each student from the student VLE dataset based on the course module and presentation. This gives an aggregated total number of interactions per course student for which to see if there is a correlation and way to predict with the total number of interactions alone.

The second dataset adds to this the area where the interactions of the students took place. This involves combining the student VLE and VLE datasets together and getting a total clicks for each activity type. This is then pivoted to get a new table for the students, course and presentation. The student information is then added to this dataset to give the second dataset analyzed.

The third and final dataset that is used in this research is the same as the second with the addition of the total scores for each of the assessment areas and then an aggregated score for the students.

C. Missing values

There are several columns that feature missing values but the only one that is used in the research which is IMD band. This is addressed by removing them completely being 1054 of 29 228 records being removed totaling 0.04% of the total data. The other features such as the week from and to in the VLE dataset are not used in this research as well as the date in the assessment dataset.

IV. Data Analysis

This section will review the analysis done on the data prior to the machine learning and predictive analytics. The aim of this section is to identify any points in the data that provide some insight and overall are interesting.

The first point to consider is the total number of clicks compared to the number grades achieved.

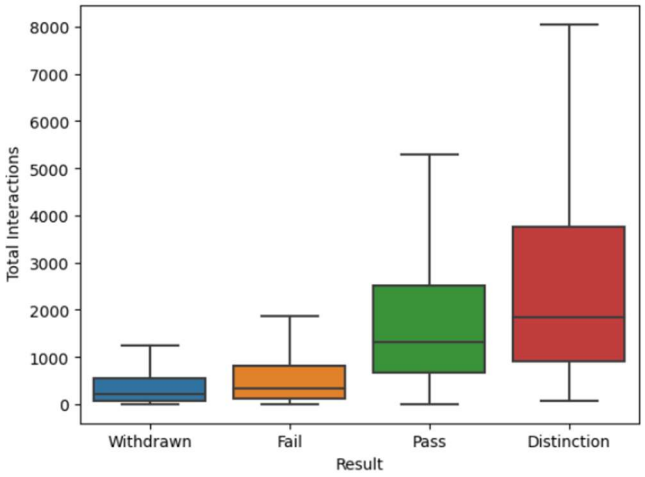

Fig. 5. Boxplot of total interactions against final result.

From figure 5 it can be seen that the more interactions the student has with the VLE the more likely they are to pass the course. The median for fail and withdrawal are close but it must be noted that withdrawal could be at the start of the course so there is less time for them to have interactions. It is interesting to know at this stage that there is definitely a correlation between success and the number of interactions.

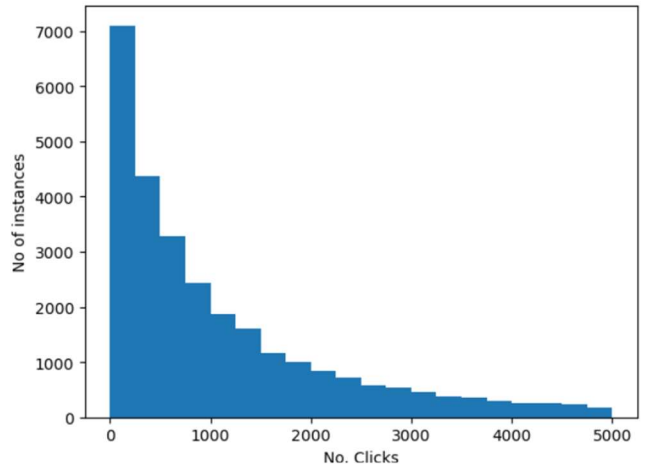

Fig. 6. Histogram of total clicks per student capped at 5000 interactions.

Figure 6 shows the number of interactions by student with a histogram and it shows an extremely skewed amount with most sitting under 2 000 interactions. It does go up to 25 000 but was reduced to 5 000 in the graph as there were very few above this level and the trend can be seen.

The breakdown of IMD banding was also of interest to see if depravity of the area impacts the performance of the students and their number of interactions that they have with the VLE.

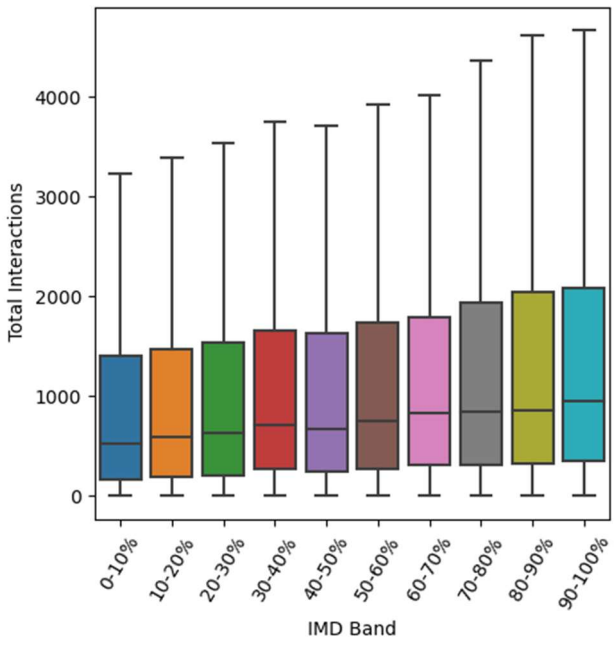

Fig. 7. Boxplot of IMD Band by Total interactions

Figure 7 shows evidence that there is a difference between the lowest and highest banding and the number of interactions that the students make with the VLE. There can of course be many reasons for this such as time, access to the online information or just a general want to undertake the course in the long run. If it is the students first course they may feel the path is not right for them and could opt for a bricks and mortar university instead. Or could have dropped out which is shown in figure 11.

When looking at the pass fail rates for the IMD bands it is clear that the lower bandings are more likely to fail and withdraw which could correlate to the aforementioned reasoning as to why there are less interactions for the students in the lower IMD bands.

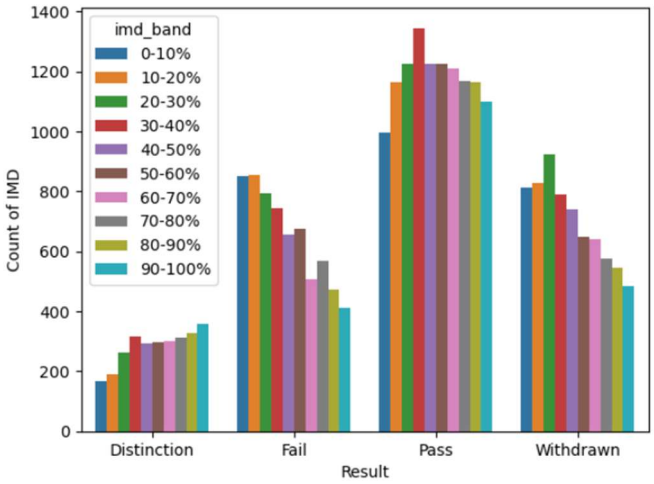

Fig. 8. Histogram of final result by IMD Band

Looking at other areas the male to female interactions has similar medians with females being lower over all and males having a bigger spread. Regions have similar medians within 200 interactions of each other. IQR for all regions is quite similar as well.

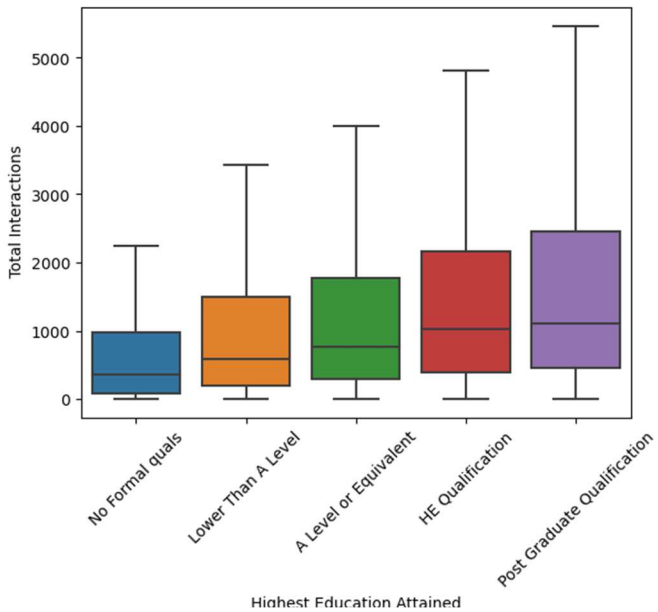

Fig. 9. Boxplot highest education level and number of interactions

Figure 9 shows that the higher the education level of the student involved the more interactions they have with the material. It has been found that students with higher education spend more time studying on average [13]. This may be down to an appreciation of the difficulty at the degree level or having already gone through the higher education process. There is a plethora of reasons for this, but nothing is evident in the data.

There is further analysis that can be done on this data such as analysis against the final results and comparing the above metrics. This is something that is noted in the future work section of this report.

V. Machine Learning Methods

This section will show the approaches and measures used with regards to machine learning. The models and metrics will be discussed. The transformation of the data and addressing issues that arose in the process.

A. Models used and reasoning

Across the 3 datasets that are to be used for the machine learning The research stuck with the same models to have a consistent comparison of performance across the additional information. The models are listed in the table below:

TABLE II. ML Models

KNN	K Nearest Neighbors
NB	Multinomial Naïve Bayes
DT	Decision Tree
RF	Random Forest
VC	Voting Classifier
GBCL	Gradient Boosted Classifier

The models are used in the above sequence for each of the 3 datasets that are analysed. The voting classifier combines KNN, NB and DT. The GBCL is used in dataset 2 but performs worse for all 4 categories as well as the 2 individual categories of pass or fail. It is added to VC as well and it drags down the performance of the VC so it is only used in one instance and then not used for the final dataset.

The KNN model starts with a base number of neighbours being 5. The optimal number is then found by looping over from 1 to 50 neighbours. Once this is found the accuracy of the model is reported and then this parameter is used for the KNN element of the VC. Across the different cuts of the 3 datasets it is evident that this is required as optimal neighbours takes many values such as 7, 21 and 49 respectively for each of the 3 datasets when looking at all 4 categories that need to be predicted.

The NB model used was the multinomial rather than the Gaussian bayes that was tested on dataset one. This resulted in a 3% increase on accuracy of the models. Although Multinomial is predominantly for NLP tasks and classification [14]. It is generally looking at frequency of appearances so this may be why it worked well in this instance.

The decision tree was again used with a default max depth of 100 and then tuned down between 1 and 100. Again the parameter was dependent on the dataset that was used. With a variety of depths such as 1, 7, 8 for different cuts of the data.

The random forest was used as the first ensemble method as a base line for what performance should be expected. It was shown in the research that the best performing model for all 3 datasets in both the 2 and 4 category problems gave the best accuracy as well as other metrics which will be listed in part B of this section.

The voting classifier was used with a combination of KNN, NB and DT. There is an instance in dataset 2 where it was tried with RF and GBCL but as mentioned above it did not perform as well as the base 3 models tuned for the dataset. The voting type used was “hard” or majority voting. “soft” argmax was also tried but the performance was lacking when compared with the “hard” voting. This is likely in the 4 category models to the accuracies being a bit lower and the models not being tuned optimally for “soft” voting.

B. Results measures

For the measure of results the following methods were used when looking at performance of the models across the 3 datasets.

- **Accuracy = TP + TN / TP+ TN+ FP+ FN**
- **Precision = TP / TP + FP**
- **Recall = TP / TP + FN**
- **F1 Score = 2* (Precision * Recall) / (Precision + Recall)**

The accuracy will be the overall correct percentage of correct predictions in the model. Describing the above will be in the context of predicting a pass. Precision will show how many predicted passes compared with the incorrect prediction of a pass. For recall it is the number of correct predictions of a pass against the number of times the model predicts a fail for a passing result. F1 score will measure the average of precision and recall and show a balance between the 2 metrics.

For the research accuracy alone is evidently not enough. When looking to aid students or mark them as at risk the model must be able to flag them efficiently. It is therefore important that precision and recall are high or at least similar and have a good F1 score. This is because incorrectly identifying at risk students could hamper their future or academic future when simple interventions could be put in place.

C. *Data transformation and imbalance*

For the datasets to be in a usable format for the models it meant transforming the data into a numerical format and making the change for any missing values. The only feature that had missing values that was used was the IMD banding, and this was addressed in section IV. To recap these were removed as there was no modal value and the performance all over the student was not something that could be predicted.

With the fields themselves some used a mapping and others used one hot encoding. The mapping was used for the IMD band, Age band, Highest education level and age. This is because it can be seen as a ranking for the different areas where the student lives, the education level and age. The other features used one hot encoding as mapping would add in bias to the models that would not be optimal for performance as items such as region are specific to the student and have no such ranking.

The final mapping was the final grades where they were mapped twice once with pass and distinction recorded as a pass and fail and withdrawn mapped as a failure. This is on the logic that the student either completed the course successfully or failed to successfully complete the course. The alternative mapping was each as 0-3.

When working with the value 2 category calculations the data was 60:40 between the pass and the fail category as seen below in figure 13. Looking at the 2022 paper by Hakim Etal. [15] it is clear that at this level there is a slight imbalance but it usable without intervention for the models. It could be addressed but would have little impact. This split is shown in figure 10.

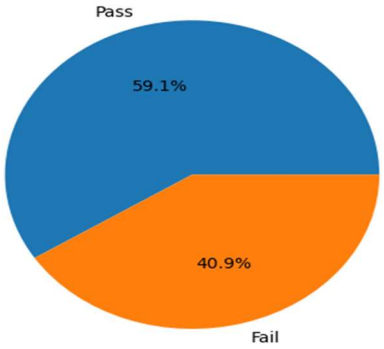

Fig. 10. Pie chart of pass and fail split for 2 categories

On the other hand, when looking at the 4-category model it is evident that the data is imbalanced and should be addressed as per figure 11.

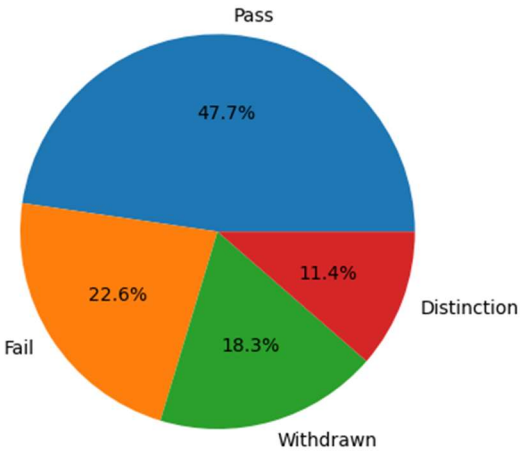

Fig. 11. Pie chart of pass, fail, withdrawn, distinction for the data set

To address the imbalance SMOTE (Synthetic Minority Oversampling Technique), Random over and under sampling were used. When measuring the performance of the models in all instances SMOTE came out on top and was therefore used for the final models. SMOTE is an oversampling method that selects minority example then using KNN creates a synthetic example between the selected item and the closest neighbor [16].

D. *Feature Selection and Normalisation*

The feature selection method used for the models was K best. An attempt was made to use Sequential feature selector both backwards and forwards it gave the full selection of features. This is likely due to incorrect hyperparameter selection when using the tool and an item for further investigation as part of future work in the project. The feature selection process came up with 9 and 10 features for the first dataset noting that the clicks, IMD band, interactions and course combinations were important features. This filters through to the second and third data sets with all the areas of interaction being key and then the grade predictions were crucial features as would be expected from the prior research. It is shown overall that the feature selection correlated with the prior research in the space showing that the best predictors were the prior performance grades, interactions and then the demographic data that is available.

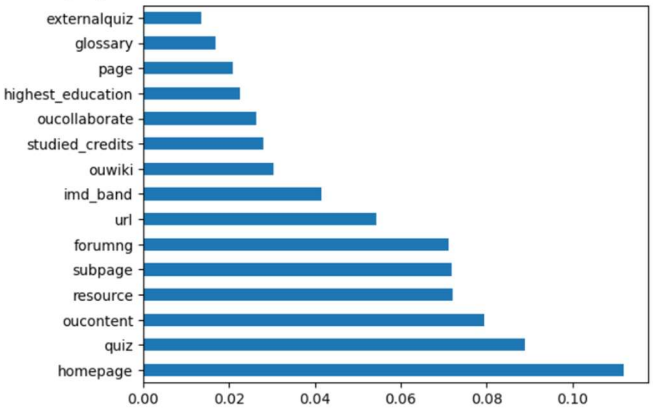

Fig. 12. Bar chart of top 15 features and their importance for predicting pass, fail, withdrawn, distinction for dataset 2.

From figure 12 it shows that the top 15 features for the random forest on dataset 2 are all the areas of interaction proving that they are key features that predict how the students will perform in the course.

Figure 14 in the appendix shows that the most significant features for dataset 3 are related to the assessments and then the next being the areas of interaction noticeably the homepage, quiz and forum interactions being the highest for the interaction areas.

Normalization was tried on the datasets, but it did not appear to increase any of the measures for any of the datasets when running the analysis. This likely be attributed to the sparsity of most of the data and the overall importance of the number of interactions, the progressing through the course and other numbered metrics are important factors in the data so normalization here was not the best course of action.

VI. Analyis of results

For this section the results for the machine learning will be described split across the 2 variants that were undertaken. This is pass or fail and pass, fail, withdrawn and distinction.

A. Results for pass, fail, withdrawn, distinction predicitons

For the report tables below, the macro average of all classes is presented. For the best performing model overall, the full breakdown of all categories and scores in each dataset will be presented. The accuracy is the overall for the model and is not an average.

TABLE III. Pass, Fail, Withdrawn, Distingction Results

Model	*Accuracy*	*precision*	*Recall*	*F1*
		Dataset 1		
KNN	42.95%	39.86%	38.69%	38.25%
DT	54.04%	**49.01%**	**50.51%**	**48.88%**
NB	40.20%	38.69%	42.62%	37.33%
RF	**55.97%**	47.44%	45.52%	45.75%
VC	45.52%	43.50%	46.58%	43.28%
		Dataset 2		
KNN	45.00%	45.34%	48.38%	43.68%
DT	55.97%	49.24%	49.79%	49.33%
NB	41.85%	38.80%	40.71%	37.29%
RF	**62.38%**	55.13%	**51.76%**	**52.30%**
VC	50.97%	**46.70%**	49.06%	47.05%
		Dataset 3		
KNN	55.56%	50.34%	52.36%	50.96%
DT	67.50%	61.25%	64.62%	61.03%
NB	39.92%	41.66%	41.94%	39.24%
RF	**72.82%**	**67.51%**	**68.00%**	**67.56%**
VC	57.64%	54.70%	60.18%	55.67%

B. Results for Pass, Fail Predictions

For the 2 categories classifiers it uses pass and fail. With distinction and withdrawn going to pass and fail respectively.

TABLE IV. Pass, Fail results

Model	*Accuracy*	*precision*	*Recall*	*F1*
		Dataset 1		
KNN	78.81%	79.13%	78.39%	78.52%
DT	82.06%	**82.73%**	81.57%	81.75%
NB	75.64%	76.36%	76.10%	75.62%
RF	**82.19%**	82.32%	**81.93%**	**82.03%**
VC	80.38%	80.38%	80.18%	80.25%
		Dataset 2		
KNN	83.39%	83.84%	83.11%	83.22%
DT	86.02%	86.48%	85.75%	85.88%
NB	73.98%	75.25%	74.45%	73.86%
RF	**87.93%**	**88.35%**	**87.69%**	**87.83%**
VC	85.00%	85.13%	84.85%	84.92%
		Dataset 3		
KNN	90.02%	90.45%	88.87%	89.48%
DT	89.48%	90.47%	87.93%	88.80%
NB	72.90%	74.04%	74.64%	72.85%
RF	**93.59%**	**93.82%**	**92.90%**	**93.30%**
VC	90.24%	90.80%	89.03%	89.70%

As can be seen from table III and IV, as the detail for interactions are added for dataset 2 and the grades then in dataset 3 it is clear that the best performing model is the random forest across all metrics. It can be noted is the more demographic based dataset 1 that a decision tree tends to perform better and the closest metric for RF to the others tends to be the recall.

C. Analysis of the RF models

1) Pass, Fail

For the 2 category model there is not a lot more that can be added other than the full table of results showing the closeness for metrics across the 2 classes. From the tables accuracy will be omitted as it is for the overall performance rather than the individual categories.

TABLE V. Pass Fail Random Forest Results.

Outcome	*precision*	*Recall*	*F1*
	Dataset 1		
Pass	81.29%	86.30%	83.72%
Fail	83.35%	77.55%	80.35%
	Dataset 2		
Pass	85.32%	92.91%	88.95%
Fail	91.39%	82.48%	86.71%
	Dataset 3		
Pass	92.84%	96.63%	94.69%
Fail	94.80%	89.17%	91.90%

From table V it is clear that even with just the areas of interactions it has a strong performance in predicting if a student will pass or fail by the metrics given. Adding in the grades and predicting would imply a strong prediction for a student could be made from the data.

2) Pass, Fail, Withdrawn, Distinction

TABLE VI. Pass, Fail, Withdrawn, Distinction Random forest results

Outcome	*precision*	*Recall*	*F1*
	Dataset 1		
Pass	64.99%	64.45%	64.72%
Fail	40.96%	37.54%	39.17%
Withdrawn	55.60%	55.48%	55.54%
Distinction	27.79%	34.38%	30.73%
	Dataset 2		
Pass	67.51%	84.92%	75.22%
Fail	51.28%	35.78%	42.15%
Withdrawn	64.16%	64.94%	64.55%
Distinction	37.58%	21.42%	27.28%

Outcome	*precision*	*Recall*	*F1*
	Dataset 1		
	Dataset 3		
Pass	83.70%	86.16%	84.91%
Fail	61.27%	50.36%	55.28%
Withdrawn	63.11%	69.48%	66.14%
Distinction	61.97%	66.00%	63.92%

Looking at table VI it is clear that for dataset 1 and 2 the model performance is not great, and it can even be argued that for dataset 3 the performance for categories other than pass are not good predictors.

Fig. 13. Confusion matrix of pass, fail, withdraw, distinction categories dataset 3 RF predictions.

From figure 13 it can be seen that misclassification is still predicting if a student will complete the course or fail to complete the course. This is demonstrated in distinctions only predicting 2 fails and 0 withdrawals and withdrawals only predicting 7 passes and no distinctions. This shows that while the measures may not have the best-looking percentages the model is still predicting accurately if it were to be abstracted to 2 classifiers.

VII. Conclusion & Future Work

This section will highlight the findings of the research and go on to describe the future work that could be undertaken for the datasets.

A. Conclusions

From this research paper it is evident that predicting student performance is doable not only with past performance of the student but how they interact with a VLE. As there is an uptick in online learning it will be a valuable measure for learning institutions to consider when assessing and monitoring the performance of students.

While the results of the pass / fail categories are not surprising as it would be expected for a model to perform well the pass, fail, distinction, withdrawn categories predictions show that metrics alone do not show the whole picture. The models are predicting students that may or may not be at risk of failure and withdrawal so could be approached with assistance to ensure that they complete their education. As figure 12 shows there are no distinction predictions for withdrawal and only 2 for fails. The vast majority of incorrect predictions are therefore only a slight deviation from the true category and not totally incorrect.

If an institution wanted to use similar data to predict the performance or risk of a student failing / withdrawing, following a similar approach to this research should enable them to somewhat accurately identify these students. It is likely possible that they could then provide additional assistance if required or identify other factors that lead to withdrawal or failure.

Outside of the machine learning and predictive parts of the research it is interesting to identify areas that indicate a higher level of interaction. The main amongst these is the shown in figure 7 where the IMD banding indicates how often the student will interact with a VLE. This could be due to factors such as not having a support network for the student, having to work longer hours to pay bills and so on. It may also just be a lack of appreciation for the requirements of education coming from a background where education was to a lower standard. This is a branch that could be followed to understand the reasoning and a further research topic.

The age and interactions are not surprising and showing male to female is similar. However, figure 9 also shows an interesting insight that the higher education level, the higher the number of interactions. This is likely due to the appreciation of the requirements and commitment needed for higher education but again another branch that could be investigated for further research.

There is analysis that is not undertaken as part of this research that is left within the data. There are many additional insights and probable factors that could be used to refine predictions down further considering unused elements of all datasets. It is for this reason that future work on this and other similar datasets could be used for ongoing performance monitoring of students and also predicting performance throughout the courses.

Overall, the research is able to answer the question that shows how often a student interacts with a VLE is a good measure of their likely success in the course. With deeper analysis referenced in further work it is likely that these predictions could become more accurate and therefore help to provide institutions with an online offering a good way to monitor and aid students that may be at risk of failure or likely to withdraw.

B. Future Work

Given the results of the investigation the future developments of this project would look to a time based approach for accessing and using the VLE. This data is available in the datasets and thus is something that can be analysed. In order to do so would be a case cutting and forming a different dataset that groups either by TMA submission date for the different courses or goes to a grouping such as interactions every 10 days to see how this impacts performance. There are a number of unexplored combinations in the data that the reader may find insightful and that may show better results than have been attained in this paper.

There is also the opportunity to explore deeper the factors that call for a student to withdraw from a course based on their interaction. There is likely to be extenuating circumstances for several of these such as bereavement, circumstances changing, economic reasoning but there may be correlation with the interactions and some students just do not get on with the course material. If further data is published then it would be possible to expand on this research to see if the same techniques can be applied to all courses or if there is a difference in requirements for the different faculties such as STEM, Humanities, law etc.

For even further analysis there is the path of looking at more complex models especially when dealing with the pass, fail, withdraw, distinction labels of classification in the dataset. This could include neural networks which are becoming more frequently used in this area as per "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance" [7] Additional ensemble models as well as different approaches to the ones used for this paper may also provide more accurate predictions. It would also be pertinent to consider multiclass classification as the results clearly show the model predicting the students in a pass / fail scenario rather accurately. It is the distinction between the classes in those areas that is lacking.

VIII. References

[1] D. Peck, "Devlin Peck," Devlin Peck, 11 January 2024. [Online]. Available: https://www.devlinpeck.com/content/online-learning-statistics. [Accessed 22 July 2024].

[2] M. L. S. R. a. A. R. Nadine Diaz-Infante, "Demand for online education is growing. Are providers ready?," Mckinsey & Company, 20 July 2022. [Online]. Available: https://www.mckinsey.com/industries/education/our-insights/demand-for-online-education-is-growing-are-providers-ready. [Accessed 22 July 2024].

[3] GTPE Communications, "7 Key Factors to Consider When Choosing an Online Degree Program," Georgia Tech, 21 June 2021. [Online]. Available: https://pe.gatech.edu/blog/working-learning/choosing-online-degree-program. [Accessed 23 July 2024].

[4] NCES (National Centre for Educational Statistics), "Fast facts," IES > NCES, 15 November 2022. [Online]. Available: https://nces.ed.gov/fastfacts/display.asp?id=1122. [Accessed 24 July 2024].

[5] G. D. S. Jacob Cornog, "PREDICTING PERFORMANCE IN CHEMISTRY," *Journal of Chemical Education,* vol. 2, no. 8, pp. 701-708, 1925.

[6] A. F. Tomas Chamorro-Premuzic, "Personality predicts academic performance: Evidence from two longitudinal university samples," *Journal of Research in Personality,* vol. 37, no. 4, pp. 319-338, 2003.

[7] Y. B. G. A. A. M. A. A. A. N. A. Yazan A. Alsariera, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," 9 May 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2022/4151487. [Accessed 26 July 2024].

[8] B. Z. A. M. Amal Alhassan, "Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data," *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 4, pp. 185-194, 2020.

[9] B. L. M. Y. K. A. Attia NA, "The potential effect of technology and distractions on undergraduate students' concentration," *Pak J Med Sci.,* vol. 33, no. 4, pp. 860-865, 2017.

[10] M. H. D. H. Z. Z. J. V. A. W. Jakub Kuzilek, "OU Analyse: analysing at-risk students at The Open University," *Learning Analytics Review,* vol. 15, no. 1, pp. 1-16, 2015.

[11] T. Devastator, "Open University Learning Analytics Dataset Student Performance and Engagement Data at The Open University," Kaggle, 20 December 2023. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/open-university-learning-analytics-dataset. [Accessed 10 june 2024].

[12] Ministry of Housing, Communities & Local Government, "The English Indices of Deprivation," UK Government - National Statistics, London, 2019.

[13] Grand Canyon University, "An Analysis of Study Habits, According to Students Across the U.S.," Grand Canyon University, 31 March 2022. [Online]. Available: https://www.gcu.edu/blog/gcu-experience/analysis-study-habits-according-students-across-us. [Accessed 30 July 2024].

[14] Great Learning Team, "Multinomial Naive Bayes Explained," Great Learning, 30 April 2024. [Online]. Available: https://www.mygreatlearning.com/blog/multinomial-naive-bayes-explained/. [Accessed 30 July 2024].

[15] M. Y. A. M. Mohd Hakim Abdul Hamid, "Survey on Highly Imbalanced Multi-class Data," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 6, pp. 211-229, 2022.

IX. Appendix

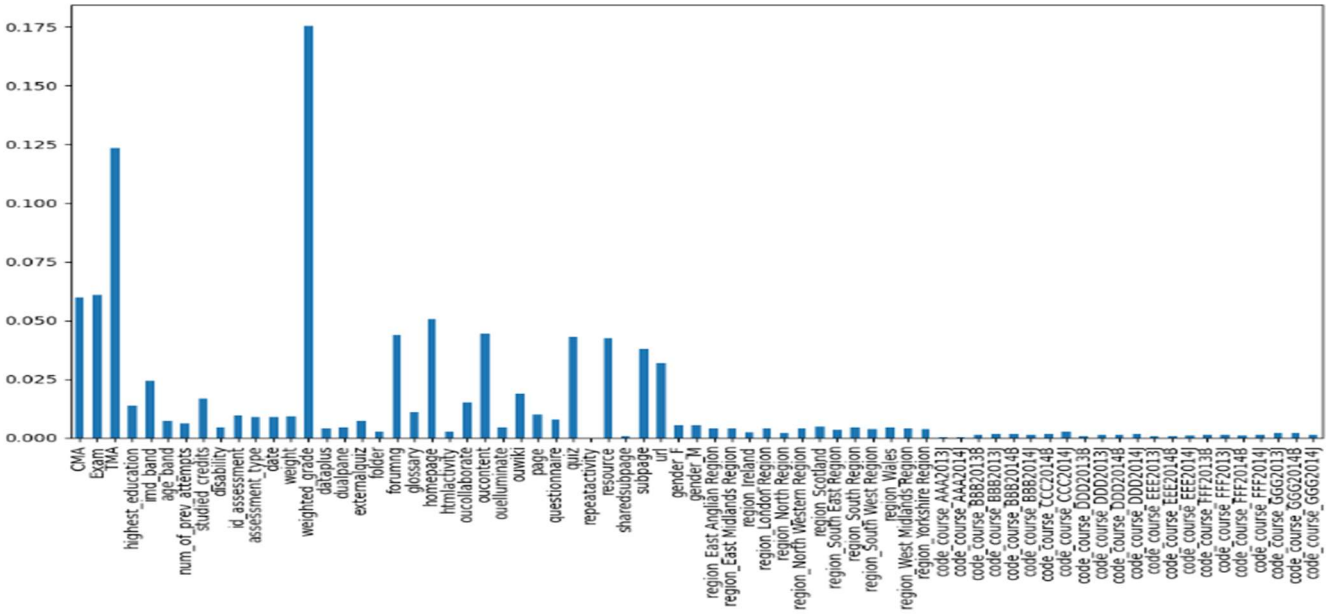

Fig. 14. Random forrest most influencial fields for dataset 3.