Research Proposal Presentation ra22895

Hello, my name is Ruth Allison and my student number ra22895. And this is my research proposal presentation for the research methods and professional practice module.

The title for my project is predicting International Foundation student outcomes based on demographic characteristics and study behaviour using classification algorithms.

I believe this project is relevant to the MSc in data science because a review of relevant literature shows that although there have been attempts to classify student outcomes, they have been with variable success, and they tend to focus on either demographic characteristics or on study behaviours, and the number of variables used tends to be quite small.

So the data set that I intend to use contains around 5000 student records. The students come from over 100 countries and this can be seen on the map on screen, with the pink areas indicating the countries where students come from. And the students arrive with more than 80 different qualifications to study in the UK.

So the question that I intend to address is can machine learning algorithms accurately categorize outcomes for international students at UK higher education institutes, based on their demographic characteristics and study behaviours.

So the reason for looking into this is that each year we have around about 5000 Foundation students that are coming to study in the UK. They come from a wide variety of backgrounds and educational systems. And these qualifications and educational systems that they come from don't align directly with the UK system, so they do a foundation course in order to then progress to their chosen University.

So they're coming from lots of different countries, lots of different experiences, different education systems, different expectations that may or may not match what is expected in the UK. And what we would like to do is to be able to classify the students so that additional support can be given. So students can be split into three broad categories. There are those who are at risk of failure. This may be because they haven't studied in the way that we study in the UK before. It may be through knowledge gaps, but we need to identify those students early. The next group of students is the ones that are likely to pass, but they may need extra support if they need to gain very high grades. And the final group is those students who are likely to gain high grades, but they may become disengaged through lack of challenge. So being able to classify students allows us to target support for weaker students and give stretch and challenge opportunities to the strongest students. And early intervention can prevent issues escalating, improving outcomes. So when a student applies to study in the UK, a large amount of demographic data is collected, but the size of the data sets and the number of variables makes analysis difficult. So this type of problem is a good fit for the capabilities of machine learning algorithms.

So the aim of my project is to create a machine learning model using classification algorithms. To predict outcomes for foundation students studying to enter higher education in the UK, and I will achieve this by testing and evaluating three different algorithms, determining which is the most accurate algorithm for the task and then training and testing the chosen algorithm with real student data.

The key literature is included on this slide. The two main papers to draw your attention to is Rastrollo-Guerror et al. This is analysing and predicting student performance by means of machine learning. This is a review of a number of papers that have attempted to classify students with varying levels of success. And it gives an overview of the different algorithms that have been used and the different feature selection methods that have been used. In terms of the algorithms, this first piece of literature Chaubey et al. Customer purchasing behaviour prediction using machine learning classification techniques. Although this is aimed at customer purchasing rather than students, there are a lot of commonalities a lot of similarity between customer behaviour and student behaviour. In both cases, there is a mixture of demographic behaviour a demographic characteristic such as geography, location, and behavioural aspects. So, there are a lot of similarities and Chaubey et al. looked at 13 different algorithms to measure the level of accuracy all on the same data set. Other pieces of literature here look at how different algorithms can be used different ways of setting them up also feature selection approaches.

So, methodology, the research is primary, we will be using demographic and study behaviour data. So, the data collected for the demographic part comes from student application forms. And here we will be collecting nationality, domicile, qualifications, age, gender, how long it has been since the student was last in full time education, and we also have information about any specific needs that the student has. For the study behaviour data, this will partly come from our student record system for this, this will give me attendance. It will also give me early test opportunity scores. So for many of our modules, we have baseline tests. These are tests that students take on arrival to determine their proficiency in a specific subject. And from our virtual learning environment, over the first two or three weeks, of their program we can also measure interactions with the virtual learning environment so we can measure how many times they've engaged with their module materials.

Once the data has been collected, it will be cleaned, anonymized and pre-processed. We will be removing incomplete records and ensuring that variables are labelled consistently. Because the majority of our variables are categorical, then we will be using a chi-squared test for feature selection. Initially three algorithms will be tested and these are naïve Bayes, ANN artificial neural network and an ensemble method as recommended by Chaubey et al., which is the KnnSgd model, where we have k-nearest neighbour and stochastic gradient descent algorithms as the base layer with a Knn meta layer.

Once these models have been sorry, once these algorithms have been tested, they'll be evaluated to determine which is the best fit for this task. And the evaluation will look at accuracy precision and recall. Once that has been done, then the most appropriate model most appropriate algorithm will be chosen to then test on the live data set.

Because we're using live data, there is ethical there are ethical considerations. So data will be anonymized. It's use It's previous academic year's data, so this will be completely anonymized. Once the data has been collated, no anonymized data will be stored and the data that is stored will be in encrypted files.

There is a risk with this type of task of stereotyping students as pass or fail. And we this will be mitigated by explicit statement for anybody looking at the results to say that this is not a definite outcome. This is a potential outcome, though this is not certain. And it should be used to support students and not pigeonhole them.

Because it is, again using live student data and there's potential risks to pigeonholing students with the outcomes, this project has gone through the ethics committee and has had input from the legal team and has received approval for me to use the data.

So ultimately, what I am intending to do is to build a machine learning model that can be used by colleagues in colleges to classify students into one of three groups at risk of failing, pass and stretch and challenge.

Ultimately, if this can be can be used by the colleges, then they can improve student outcomes by offering support to weaker students and continuing to engage stronger students with stretch and challenge activities, thus, improving outcomes across the board.

And finally, here is the timeline for my proposed project. As you can see, the project has been split into three phases. Phase one is data collection and pre-processing. So this will be collecting the data, creating the dataset, combine combining all of the data together, cleaning to remove incomplete records and pre-processing to convert categorical data to factors, split into test and train, and to anonymize the data.

Phase two is model testing and evaluation. So the three different algorithms will be tested evaluated, and then the most appropriate model, the most accurate one for the task will be selected for further testing and evaluation. Then finally, there will be the project write-up.

Thank you very much.