The Prediction of Student Performance Through the Use of Machine Learning

Alasdair Bruce

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Software Development

Dept. of Computer and Information Sciences

University of Strathclyde

August 2019

DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself.

Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

(please tick) Yes [✓] No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices is 16619

I confirm that I wish this to be assessed as a Type 1 2 3 4 (5) Dissertation (please circle)

Signature: 

Date: 18/08/2019

ABSTRACT

With the increasing popularity and importance of Higher Education, degree cohorts are becoming larger and it is becoming increasingly difficult for lecturers and advisors of studies to engage with students personally. As a result, students who are struggling with course material are receiving feedback and advice too late or are being missed out altogether which increases the number of students failing or dropping out of higher education.

The increase in research into machine learning and advances in technology have made complex and computationally expensive machine learning algorithms a viable solution to this problem. The aim of this study was to determine which machine learning algorithms were most suited to solving this problem and to determine whether it is possible to accurately predict a student's performance early in a course.

Using Scikit Learn, five machine learning algorithms were trained and tested using the Open Universities Learning Analytics data. The percentage of correctly classified failing students to the total number of failing students in the testing data set and the percentage of failing students correctly classified to the total number of students classified as failing were used as metrics to determine which of the algorithms were most suited to this problem.

The results show that, with minimal amounts of parameter tuning for optimisation, Random Forest and K-Nearest Neighbours are both suited to predicting student performance even with only a small amount of available student data which would allow for accurate and early prediction.

Acknowledgements

I would like to thank my dissertation supervisor Konstantinos Liaskos whose help throughout the entire MSc programme has been invaluable, not just for myself but for the entire cohort.

I would like to thank my parents for being supportive of my decision to return to university and for their encouragement over the past year.

Finally, I would like to thank my partner for offering her support and pushing me to apply for this degree. Without that push I wouldn't be where I am today.

Contents

1. Introduction

1.1 Background and Motivation

Higher education is increasingly seen as essential by many employers. This has led to an increase, not only in the number of people opting to pursue degrees in all fields, but also in the diversity of the students. While in the past students were likely to be affluent, young males seeking an academic career in a specific field, the backgrounds and motivations of students today are significantly more diverse. Due to the increase in student numbers and diversity it can be difficult for an advisor of studies to get to know each student personally and provide meaningful feedback on their academic performance. This can ultimately lead to student dissatisfaction and an increase in the number of students failing to complete their degree on time or dropping out of university.

By identifying students at risk of failing early and providing feedback at an early stage in a degree it is believed that student performance can be improved and student retention rates can be increased, outcomes which benefit both the students and the institutions. A tool or application, made available to students and lecturers, which can take in student information and provide information of their expected performance and provide feedback on how to improve their current situation would be greatly beneficial in this regard.

Technological advancements in recent years have made computationally expensive prediction methods such as machine learning a viable option for a variety of applications such as this and, with a great deal of research having gone into the development and improvement of machine learning algorithms, there are several algorithms available which are suitable for this specific task. Identifying a suitable algorithm is the first step in developing an application which can be used as means of properly engaging with students who would benefit most from engagement from lecturers.

1.2. Project Aim

The focus of this project will be to compare the performances of five machine learning algorithms to determine which are most suited to the task of predicting whether a student is likely to pass or fail a course. To be considered suitable, a trained machine learning algorithm must meet two criteria. First it must be able to correctly identify failing students from existing data sets of past student results, misclassifications of failing students as passing students must be kept to a minimum. Secondary to this, misclassifications of passing students as failing students must also be kept to a minimum. While the first of the two criteria is the most important, both must be met for an algorithm to be considered suitable.

The greatest positive effect on a struggling student will be gained by identifying and engaging with them as early as possible. As such, another objective of this study will be to investigate how much relevant student grade data is required to produce accurate predictions and to determine if any other available student data can aid in providing improvements to the performance of the models.

The project aims can therefore be summarised as follows:

- To determine if machine learning models are viable as a means of identifying students at risk of failing.
- To determine how early in a course it is possible to accurately identify whether a student is likely to fail or not.
- To compare a range of machine learning algorithms and identify those which may be suitable for this purpose.

- To determine how differences in data sets affect the performance of machine learning algorithms in an academic context.
- To determine whether features other than student grades are beneficial in making predictions.

2. Literature Review

This section will be split into four main sections, with each part highlighting previous works relevant to that particular area. The first section will discuss studies into the correlation between grades achieved earlier in a student's career, such as mid-term assessments or the early years in a multi-year course, and the final outcome of a class or course. While it may seem obvious that early grades will have a direct correlation with later grades it is important to develop an understanding of a variety of related factors such as what kinds of assessment correlate most, what is the earliest grade which can be used as a predictor before correlation drops off and how closely must subjects relate to act as valid predictors.

The second section will explore how a student's background factors into a student's performance throughout their education. Factors such as gender, upbringing, parental education and level of poverty need to be investigated as it is not immediately clear how, or even if, these factors will have any effect and as such will allow for easier identification of relevant data pre-processing data.

The third section will discuss previous studies offering information on the best performing and most reliable machine learning models and the methods available for improving the performance of those models. It will also discuss how to determine which models are best suited for any particular situation and what factors most affect model performance. By identifying the strengths and weaknesses of each model it will be possible to narrow down those available and maximise the performance of each.

The final section will discuss the justification for feature selection, the feature selection methods available and the suitability of implementing feature selection methods in a given situation. This is another factor which allow for maximisation of the performance of the selected machine learning models.

2.1 Student Grades as Predictors

There have been many studies, especially in recent years, which have explored the relationships between the various teaching methods employed by educators and their effects on the grades of students. Many of these studies focus purely on the level of improvement seen through the use of particular assessment methods such as mid-term quizzes, practical assessments and written coursework while others investigate the impact of different kinds and levels of feedback. Previous grades in a class are expected to hold the greatest weight in predicting a student's expected outcome, especially in this context where grades will be taken from mid-term assessments which should indicate a student's early grasp of a subject.

Day et al discussed the effectiveness of exam style assessments, written assignments and mid-term quizzes as well as the effect of lecturer feedback on a student's performance (Day, Blankstein, Westenberg, & Admiraal, 2018). Through this study it was found that there is a clear correlation between assessments taken by students and the final grade of a class. A study carried out by Zhang and Henderson exploring the usefulness of formative assessments in improving and predicting student exam performance also posits that it may be possible to identify students likely to perform poorly in exams based on performance in the assessments (Zhang & Henderson, 2015). From these studies it becomes clear that using a student's grades it should be possible to determine, with reasonable accuracy, how they are likely to perform in future assessments.

Day et al also found that providing assessment feedback to most students often results in a slight overall improvement in final grades but identifying students who are at risk of failing and providing more detailed, corrective feedback may be required. Lemus-Zúñiga et al agree with these findings (Lemus-Zúñiga, et al., 2015). The results of this study show that monitoring student progress allows

teachers to determine the level of feedback required for each student and give more focussed feedback to those that need it as assessments are complete. The study also provided further evidence that providing direct, detailed feedback to underperforming students can significantly improve the performance of those students, this is especially true when the feedback is returned in a timely fashion. As such, it is clearly important that any at risk students be identified early in a course to ensure that corrective feedback can be provided as soon as possible.

Shaw and Bailey carried out a case study on the predictive validity of earlier education performance for use in determining the higher education performance of prospective students (Shaw & Bailey, 2011). High school SAT scores and early university grades were taken into account as possible predictor. The effect of advance placement subject was also investigated to see if there was any impact on student performance. This study found that there was a direct correlation between all three of these factors and a student's overall performance in university. Of note, it was also shown that there is a positive correlation between students taking relevant advanced placement courses prior to university and their final outcome when compared with students with similar SAT scores.

A study by Adamson and Clifford looked at the correlation between grades earned prior to a higher education and outcomes of the first three years of an engineering degree and a final BEng project in two different universities (Adamson & Clifford, 2002). For one of the universities an MEng project is also checked, however this is shown to be have poor correlations, possibly due to inconsistent project structures. Three A-level grades were considered as predictors, one of which was Mathematics, the other two were the best two A-levels available for each student. A prior knowledge assessment, taken by students entering their first year of study was also considered as a possible predictor and the correlation between the PKA and each of the outcomes were investigated. Finally, the correlation between the outcomes of the years were also checked with each other and the final project. The results of the A-levels rankings show medium to strong, positive correlations between A-level grades and the PKA, first, second and third year results, with the correlation dropping in the later years. There are weak positive and negative correlations with the BEng projects suggesting early grades are not good indicators of a student's more practical abilities. The PKA does have a positive correlation with all three years but the degree of correlation differs significantly between universities. Similar to A-levels, the PKA does not seem to be a good predictor of project results as the correlations are almost non-existent. All university years show a strong positive correlation with each other and with the BEng projects. This supports the hypothesis that results taken earlier in a course will be good predictors of outcomes later in a course.

Birch and Rienties carried out a similar study focussed primarily on students entering an engineering based degree looking at how A-level grades in classes related to the degree, such as Physics and Mathematics, affect individual course module grades and the average grades achieved in the first one or two years of higher education (Birch & Rienties, 2014). The results of this study show that there is a clear positive correlation between the earlier results and results gained in the first year of higher education and a lower positive correlation with second year indicating that grades in earlier relevant studies can be considered a valid predictor for a limited time period.

2.2 Student Backgrounds as Predictors

Along with studies into how previous grades affect student performance there have also been a significant number of studies into the relationship between degree exam performance and a student's general background in an effort to determine which factors have a direct effect on the outcome of a student's education. Identifying whether or not there is a relationship and to what degree that relationship has an effect on the outcome would increase an educators ability to become familiar with

a student's background and be prepared to provide more meaningful advice for each individual student which can aid in reducing student dissatisfaction, increasing student retention rate and ultimately improving student grades.

Chee et al conducted a study to whether environmental factors have a noticeable effect on each gender's performance (Chee, Pino, & Smith, 2005). The data used in this study included GPA, parental education, race, SAT score, time spent studying and a breakdown of time spent on other activities such as participating in societies and carrying out volunteer work.

The results of this study show that the females with strong academic ethics perform slightly better than males with similarly strong academic ethic. It is also shown that male and female students are affected differently by environmental factors for example male academic performance is affected more by their employment status, class attendance while females are affected more by race, parental education and academic ethic. This shows that there is a correlation between gender, however that correlation is affected by other factors which may affect the validity of gender as a predictor.

A study carried out by Tieben and Wolbers (Tieben & Wolbers, May 2010) which investigated the socio-economic impact of a student's upbringing on the eventual outcome of their education noted that students with more privileged backgrounds achieved higher results or possessed skills which prove advantageous in an academic environment. This suggests that poverty levels and possibly even the area in which a student lives could affect, and therefore aid in predicting, how a student is likely to perform throughout their academic career.

A study carried out in Serbia by Teodorovic in 2011 looked at several aspects of students such as gender, ethnicity and school background to determine which contributed most to student performance between two classes, mathematics and Serbian language (Teodorovic, 2012). From the results it was found that student affluence was one of the biggest factors to perform significantly more poorly in mathematics-based classes while grades in Serbian language were less affected, this smaller effect is believed to be due to everyday use of the language while mathematics may not be used often outside of school. Gender had a very small effect on grades with females performing better in Serbian language and males performing better in mathematics. This is likely due to traditional bias pushing students to perform better in specific subjects based on their gender, a factor which should be less prevalent in a higher education environment as preferences for subject should be well established in a student. This does suggest that in higher education gender is likely to have less impact as gender bias is expected to be less prevalent.

2.3 Machine Learning Algorithms

In the last few years machine learning has become much more popular as a tool for data analysis and prediction. As a result, there have been many studies looking into developing new machine learning algorithms and studying how to determine the best machine learning models for a whole host of different situations. Machine learning has also been suggested as a proposed tool for improving student experiences and to determine which factors in a student's university career most significantly affect the ability to predict that student's performance. The problem to be solved in this case is one where a machine learning algorithm needs to determine whether a student will pass or fail based on a number of input variables, referred to as features. In more complex problems the algorithm would be required to determine what grade or even degree outcome a student is expected to gain. All of these cases are supervised machine learning problems which require a classification model to solve.

Tan and Gilbert investigated how to determine the best machine learning model for any particular purpose and what kind of kind of algorithm typically performed best (Tan & Gilbert, 2003). A mixture

of rule-based (Decision Tree, One Rule and Decision Rules), statistical (Naïve Bayes, Instance Based, Support Vector Machines (SVM) and Neural Network) and ensemble (Stacking Bagging and Boosting) machine learning models, used as binary classifiers, were compared as part of this study along with four sets of unrelated data of varying size. In order to evaluate each outcome confusion matrices were used to calculate the accuracy and the positive predictive accuracy of each model used on each data set. The results of this study show that the performances of the models examined are highly dependent on the available data and the desired outcomes. When considering the features of the data sets, statistical models perform better with continuous features while rule-based models perform better with discrete features. When considering the information needed from predictions based on the data used it is given that rule-based models provide simpler, more understandable outcomes than other models. Finally, it is shown that ensemble models tend to perform better than any individual model. It is hypothesised that, when only small data sets are available ensemble methods may provide a better prediction than any individual model by averaging outcomes. From this it may be that ensemble methods are capable of adapting to the variations which will be found in real student data and may give the most consistently strong results.

Maclin and Optiz carried out a study evaluating the performance of Bagging (Bootstrap Aggregating) and Boosting ensemble machine learning methods using Neural Networks and Decision Trees (Maclin & Opitz, 1997). Neural Networks were tested using a single Neural Network and ensemble Neural Networks using a simple ensemble method, Bagging, Arcing Boosting and Ada Boosting methods. Decision Trees were tested using a single Decision Tree and ensemble Decision Trees using Bagging and Ada Boosting. All models were evaluated using error rates after training and testing on several unrelated data sets. In almost all cases, the simple and Bagging ensemble methods reduced the error rate by a significant amount compared to an individual model. The Boosting methods had much more varied results depending on the data set being used, in some cases the improvements were significant but most of the improvements were less than with the Bagging method. As such, it appears Bagging is a much more reliable ensemble method, especially when using varying data sets. When comparing ensemble Neural Nets and end ensemble Decision Trees no one method stood out as better than the other. Bagging and Boosting were also shown to improve both models in a similar way showing that the improvement is dependent on the data set used and not on the model.

Jacob et al carried out a comparison of a variety of classification models known for having relatively high accuracy scores using a selection of multiple performance characteristics (Jacob, Sridhar, & Murugavel, 2017). The models compared were Bayesian Network, Sequential Minimal Optimisation (SMO), J48, Random Forest, Logistic Regression and Multilayer Perceptron. The performance metrics used for all models were time taken to train the model and the accuracy of the model which was calculated as the percentage of correct predictions made in a test data set. This study also looked at how correlation-based feature selection affected the performance of each model. From the results it can be seen that the SMO model has the highest accuracy, usually by a significant margin, on all data sets used for this study, with Logistic Regression and Random Forest performing second and third in terms of accuracy. All other models varied in their performance with respect to each other. While Logistic Regression and SMO performed better than Random Forest in terms of accuracy, in most cases they performed worse in the time taken by a significant amount which would be significantly more problematic on larger data sets as the accuracy is unlikely to rise by much, especially for more accurate models, while the time taken is likely to rise linearly with the amount of data available. The results of all models after feature selection show no significant changes to accuracy for any of the data sets, however almost all results show a reduction to the time taken. The correlations between features in each data set were not given as part of this study so it is not possible to determine the level of

correlation between the features and the outcome in order to determine which features may have had the most impact on the accuracy when removed.

A similar performance study by Amancio et al compared 9 different classification models which were: Naïve Bayes, Bayesian Network, C4.5 Decision Tree, Random Forest, Simple Classification and Regression Tree, k-Nearest Neighbours, Logistic, Multilayer Perceptron and Support Vector Machine (Amancio, et al., 2014). In this study the performances of the models were measured using the accuracies of each model while varying the number of features, the number of classes and the number of elements for each class on a series of artificial data sets. A comparison was also made between the accuracies of models using the default parameters in the library used and models with varied parameters. While comparing the default parameters, the accuracies of two individual models across all data sets and data configurations gives a clear indication of the best performing defaults models. K-Nearest Neighbours performs better than all other models with the Perceptron performing better than all but the KNN model, however it is noted that KNN performs significantly worse with less features than many of the other models. Naïve Bayes performs well with a low number of features compared to others and Random Forest performs almost as well as the Perceptron model with a higher number of features.

Anderson and Anderson conducted a study in 2017 attempting to use machine for student grade prediction compared Naïve Bayes, K-Nearest Neighbours and Support Vector Machine algorithms (Anderson & Anderson, 2017). This study used historical grade data only from previous years as a data set with 10-fold cross validation to split the data and evaluate performance. In this study it was found that SVM outperformed the other classifiers used but it is noted that there is a significantly higher computational cost in using SVM over simpler methods, which may not be worth it for a marginal performance gain. It is also noted that lower performance using the simpler models may be caused by using only grade data results as features which have a high correlation between each other, a factor known to negatively affect the Naïve Bayes classifier.

2.4 Feature Importance and Selection

Given that the suitability and performance of any machine learning model is highly dependent on the data set being used, it is important to ensure the selection of features from the data set is optimised to maximise the performance of the selected model. As with the algorithms themselves, there are many studies on feature importance and several methods available for determining which features to use in a given context.

Kira and Rendell proposed that reducing the feature set would increase the speed of training by reducing the amount of data being processed, improve the quality and relevance of the data being used and increase the overall accuracy of the model (Kira & Rendell, 1992). This hypothesis was supported by Khalid et al in a separate study in which some of the more popular feature selection and feature extraction techniques were analysed in order to determine if they were indeed effective in improving the performance of machine learning models (Khalid, Khalil, & Nasreen, 2014).

In a study comparing wrapper and filter methods of determining a useful subset of data from the available data Hall and Smith hypothesise that “Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.” (Hall & Smith, 1999) The wrapper method in this case was a ten-fold cross validation of the data being used to train while the filter method used was correlation-based feature selection, the accuracy changes were checked using Naïve Bayes and Decision Trees. The results of this experiment showed that correlation-based feature selection improved accuracy on both models more reliably when there were less than twenty features in the data set while the wrapper method improved accuracy when there

were more features. Both methods did sometimes result in degradation of accuracy so it is important to ensure testing is carried out with all features before feature selection to make sure that it was necessary.

Forman found that the SVM model tends to have a high performance with all available features and that feature selection may be unnecessary. It was also found that decision tree-based models may benefit from feature selection when any class is over represented in the data set being used (Forman, 2003). This is particularly interesting as it is likely that, in much past degree data, the classes will be skewed in favour of passing students.

Liu et al investigated how heavily skewed data sets affect feature selection (Liu, Kubler, & Yu, 2014). It is shown in this study that there is no definitive best method of feature selection and the effectiveness ultimately depends on the machine learning model, the data set and the desired outcome, it was also found that, for data sets with large numbers of features, feature selection will only make any meaningful positive difference if the distribution of classes is balanced. This may not be the case in data sets with smaller numbers of features even when the classes are skewed.

3. Methodology

This section will describe the tools used to develop an environment in which data sets can be used to train and test machine learning models and provide details of the results of testing.

3.1 Requirements

The application will be a Python program which will take in a data set and process it into a format useable by Scikit Learn's algorithms. The data will then be split into training and testing data sets. The model will be trained using the testing data set and performance results will be obtained using the results of the testing data set.

The final requirements for the proposed application are as follows:

- Data sets must be read in as a single CSV file and converted into a format which is compatible with the machine learning library used.
- It must be possible to remove any features from the data as desired.
- Once the data has been processed the environment must be able to use the available data to train the model and then test the trained models performance.
- Each model must be able to be tested independently of any others.
- The percentage of correctly identified failing students out of the total number of failing students in the test data set must be provided as a metric.
- The percentage of correctly identified failing students out of the total number of students classified by the model as failing must be provided as a metric.

3.2 Build

3.2.1 Technology Used

This section provides a short description of the technologies used in developing an environment in which the data could be processed, and the machine learning models could be implemented and evaluated.

Python 3.6.5

Python is a high-level, object-oriented programming language. Python is generally considered an ideal language for developing high quality applications in short periods of time due to it being a relatively concise language compared to other languages such as Java and for having a simple syntax which makes its code very readable and easy to learn. Due to its popularity Python also has access to many large, well-maintained libraries which give access to a great deal of essential functionality. Amongst these are libraries specifically designed to aid in the development of AI and Machine Learning applications.

PyCharm IDE 2018.2.3

A good IDE can provide all the functionality required to develop an application as efficiently as possible. PyCharm is a professional level IDE which provides quality of life functionality such as easy refactoring, code completion and debugging tools.

Scikit Learn 0.21.2

Scikit Learn is a Python library[1] specifically created for the development of machine learning applications. Scikit Learn comes equipped with a large number of machine learning models suitable for different kinds of machine learning problems, including regression problems and classification problems, which due to the popularity of the library, are regularly maintained by experts in the

[1] https://scikit-learn.org/stable/

machine learning field. The library also contains several methods for evaluating each model which allows for easy analysis of results.

Pandas 0.24.2

Pandas is a Python library[2] useful for creating structured data sets and for data analysis. Pandas is perfectly suited for reading in data from a CSV and organising it into a data frame which can be used by Python. Pandas also has functionality to detect null values and drop any rows which contain null values by using the dropna() function which is useful when using large external data sets as it cannot be guaranteed that no errors have occurred and there is a value for every point of data. Failing to correctly remove all null data in a data set would result in the algorithm failing to run. For each course all of the data comes from a single CSV file, with Pandas, splitting this data into features and labels is also a trivial task.

NumPy 1.16.4

NumPy is a Python library[3] which provides additional functionality for mathematical operations on arrays. In the context of this project it is necessary for the creation of n-dimensional arrays which are necessary for the algorithms provided by Scikit Learn to process the data.

3.3 Data Description

The data for this experiment came from the Open University Learning Analytics data set[4] which contains anonymised data from single semester courses including the personal details, assessment marks achieved throughout the course and whether a student passed, failed or withdrew from their chosen course.

Each available course was given a three-character identifier which anonymised the course name and degree that the course belonged to, so it was impossible to know if any of the courses were in any way related in terms of degree or even subject matter. A list of assessments for each course was provided showing the assessment identifier code, the type of assessment and the weighting of each assessment. The types of assessment provided were tutor marked assessments, for which students received a percentage mark ranging from 0 to 100, and computer marked assessments, which were multiple choice assessments with five questions were marked out of 100 in steps of 20. The personal information provided for each student consists of a unique student identifier, the student gender, the region in the UK the student comes from, the highest qualification achieved previously, the IMD (index of multiple deprivation) band of the student, the age group the student belongs to and the students final course result. The individual marks for each assessment for all students were also provided.

From the available courses, those used in this study are AAA and BBB, referred to as Course A and Course B respectively in this paper. Course A consists of five tutor marked assessments and no computer marked assessments. Before pre-processing there are 748 student examples available. Course B consists of six tutor marked assessments and five computer marked assessments. Before pre-processing there are 5572 student examples available.

These courses were selected to the significant differences in the number of examples available, the number of features available and the relative skew in passes to fails which would allow for the identification of how well each algorithm performs with realistically varied data as it is unlikely in reality that all degree programmes will have the same level of available past data to draw upon.

---

[2] https://pandas.pydata.org/getpandas.html
[3] https://www.numpy.org/
[4] https://analyse.kmi.open.ac.uk/open_dataset#data

Due to the format large portions of the data was provided in, it was necessary to carry out some pre-processing to convert it to a numeric form which could then be used in the machine learning models. The assessment mark data was provided in numeric form and required no alteration. The personal information of the students and the final outcomes were given in text format so numbers were assigned and the data was transformed. Table 1 to Table 6 shows the original form the of the data and the numerical assignment given after pre-processing of the data.

Gender	
M	1
F	2

Table 1: Gender Pre-Processing

Student Region	
East Anglia	1
East Midlands	2
Ireland	3
London Region	4
North Region	5
North Western Region	6
Scotland	7
South East Region	8
South Region	9
South West Region	10
Wales	11
West Midlands Region	12
Yorkshire Region	13

Table 2: Student Region Pre-Processing

Highest Previous Education	
Lower than A Level	1
A Level or Equivalent	2
Higher Education Qualification	3
Post Graduate Qualification	4

Table 3: Highest Previous Education Pre-Processing

IMD Band	
0-10%	1
10-20%	2
20-30%	3
30-40%	4
40-50%	5
50-60%	6
60-70%	7
70-80%	8
80-90%	9
90-100%	10

Table 4: IMD Band Pre-Processing

Age Band	
0-35	1
35-55	2
55+	3

Table 5: Age Band Pre-Processing

Final Result	
Withdraw	0
Fail	1
Pass	2
Distinction	3

Table 6: Final Result Pre-Processing

After pre-processing, the data was combined into a single CSV file which could be read by Python. The CSV file was read in and converted into a single Pandas dataframe, student examples for which the final result was a withdrawal were then removed as leaving them in resulted in outliers as often the students grades in all assessments were zero, all student examples for which the final result was a distinction were treated simply as passes as, for the purposes of this study, it is not necessary to know how successfully a student passes, it is only important to know that they pass. This also simplifies the classification as it will be a binary classification problem.

The next step was to remove all examples containing any null values from the data as attempting to use null values in any machine learning algorithms resulted in errors.

3.4 Model Selection

As mentioned earlier in this report, there have been many studies into the most effective classification models available for prediction and it is clear that this is dependent on the data set available and the purpose. As such, multiple models will be selected for comparison in order to determine if a suitable candidate can be identified. This section will describe each of the models selected and offer a rationale for their selection.

3.4.1 Gaussian Naïve Bayes

The Gaussian Naïve Bayes model is a relatively simple machine learning model based on Bayes theorem which used for solving classification problems. Bayes Theorem can be shown by the equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where P(A|B) is the probability of outcome A while condition B is true. This is a popular model due to the relative simplicity in setting up and running this algorithm and the speed with the overall speed in predicting outcomes. The Naïve Bayes model performs well when the data set available has a very small number of features and when the data set itself is small, in these situations Naïve Bayes may even outperform more complex models, especially using the default Scikit Learn parameters. Given the data sets being used in this experiment will not have a feature quantity exceeding 20 and the available data records will not exceed ten thousand which is idea for helping to increase the performance of this particular model.

Unlike many other more complex models, Gaussian Naïve Bayes does lack tuning parameters and, while this does contribute to the simplicity of the model, it also means that any improvements to accuracy will have to be gained through data processing.

Another potential issue is that the Naïve Bayes algorithm assumes that all features in a data set are fully independent of each other. As such, the performance of Naïve Bayes may be adversely affected due to the correlations between marks achieved in assessments. Conversely this may prove to not have any negative impact when using a smaller selection of grades as features.

As a simple algorithm which relies entirely on the data set for its performance Gaussian Naïve Bayes is an important algorithm to explore as it will provide insight into the degree to which differences in the size of the data set and the available features affects all algorithms.

3.4.2 K-Nearest Neighbours

K-Nearest Neighbours is one of the simplest and most popular tuneable algorithms used to solve classification problems (Cunningham & Delany, 2007). Classification of an example is carried out by plotting the example to be classified against all the training data and determining the nearest training data set examples, a weight can then be assigned to each of these neighbours which determines how influential each training example is – this weight is usually determined by the distance to the neighbour – or each neighbour can be considered to have an equal weight regardless of distance, a majority vote on each of the nearest neighbours will then determine the predicted class. K-Nearest Neighbours is considered an instance-based algorithm as no training is carried out prior to making a prediction, instead the training data is stored in a database and queried when a prediction is required. This reduces the training time but can result in longer prediction times, especially when using data sets with large numbers of features or with many training examples.

One of K-Nearest Neighbours' most important parameters for use in tuning is *k*, which is the number of neighbours which will be identified and used when attempting to classify an example. By increasing *k* the variance will be reduced while the bias is increased which will reduce the complexity of the model and reduce overfitting, however increasing *k* by too much will result in underfitting due to the increase in bias so it is important to find a value for *k* which does not increase bias too much. The best value of *k* will differ from data set to data set and will require investigation to determine.

Figure 1 shows an example of a prediction being carried out. In this example *k* is equal to five which will result in the five nearest neighbours being considered. It can be seen that of these five, four are in Class 1 while only one is in Class 2. Assuming in this case that the weighting is uniform, the majority vote will be Class 1.

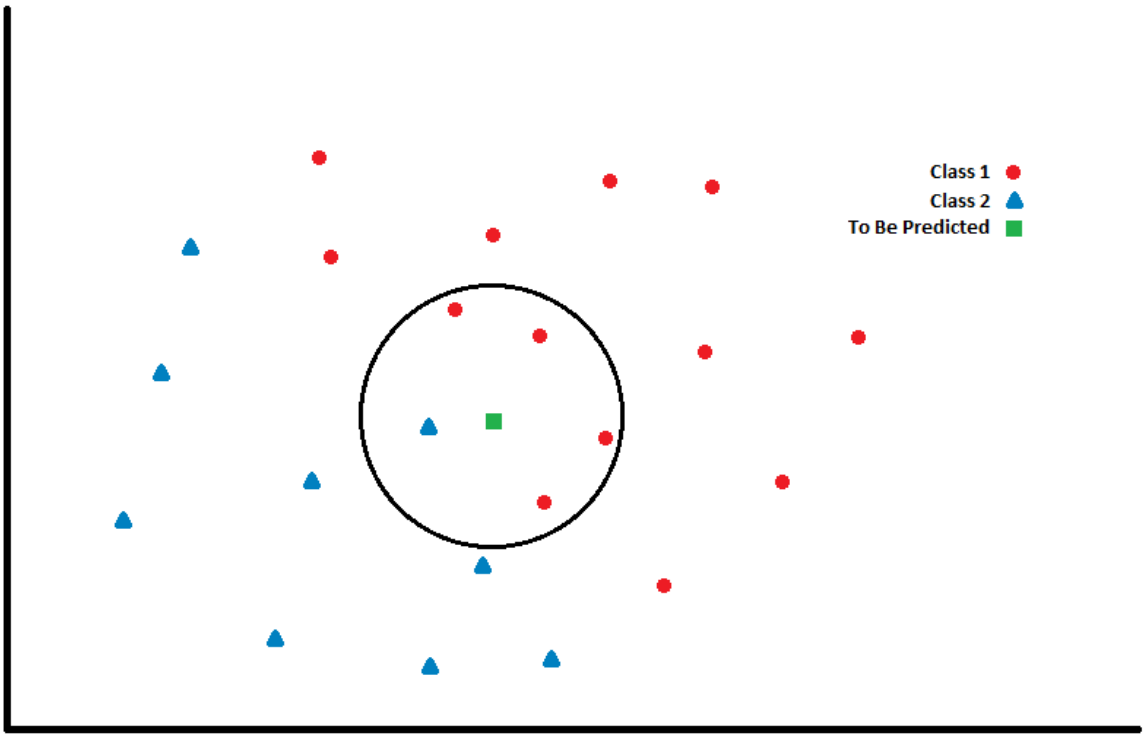

Figure 1. K-Nearest Neighbour Example

K-Nearest Neighbours tends to work best with features which are continuous as the geographical location of each example will tend to vary more. This is ideal for grade data where a percentage mark is given but is likely to be more problematic for examples containing discrete features such as gender. For this reason, it will be important to remain aware of the effect student background features have on the performance of the model.

3.4.3 Random Forest

The Random Forest model is an ensemble machine learning model which consists of multiple decision trees where the trees are built using random samples of training data and random subsets of features to determine node splits.

A decision tree is itself a machine learning model which is built by taking the training data set, splitting it down and building a tree of nodes where each node splits depending on the feature conditions. Every feature used is considered a parent node and the data the parent splits into are child nodes, which themselves split into further child nodes until a classification is decided. This is a very simple model for humans to understand as the model resembles a flow chart when visualised. They are also very fast to train when compared with other classifiers.

Decision trees have their weaknesses however, for example, allowing a decision tree to split each feature into the smallest possible pieces of data will result in massive variance increases and overfitting, while pruning of data to limit the overfitting can increase bias and end up reducing the accuracy considerably. Changes to the training data set can also result in significant changes to the accuracy of the model.

The Random Forest model overcomes the weaknesses of a single decision tree by combining several decision trees, with no depth limit, into a single model. Each individual tree will have high variance due to the lack of depth limit, but by varying the training data samples used and the features used to

determine node splits and then, when it comes to testing, averaging the predictions of each tree the overall variance can be reduced. In this way the variance problem is dealt with which prevents overfitting while any increases to bias are limited. For the available data this makes Random Forest ideal as there are cases where the data will be heavily skewed which will introduce more bias. By reducing any additions in bias the model may be able to balance variance and bias and maintain high performance.

3.4.4 Support Vector Machine

SVM (Support Vector Machine) is a machine learning algorithm which is more complex than the likes of Naïve Bayes and K-Nearest Neighbours but is no less popular. The early stages of classification with SVM are similar to K-Nearest Neighbour in that the training examples are initially plotted in an n-dimensional plot where n is the number of available features in the data set. The SVM algorithm then determines a linear (n-1)-dimensional hyperplane, or decision boundary, which splits the training data in such a way that examples from each class are on separate sides of the decision boundary while maximising the distance between the boundary and the nearest points from each class. Figure 2 shows an example of a binary classification problem where the maximisation of the margin between the nearest examples has been sacrificed in order to prevent any classification errors.

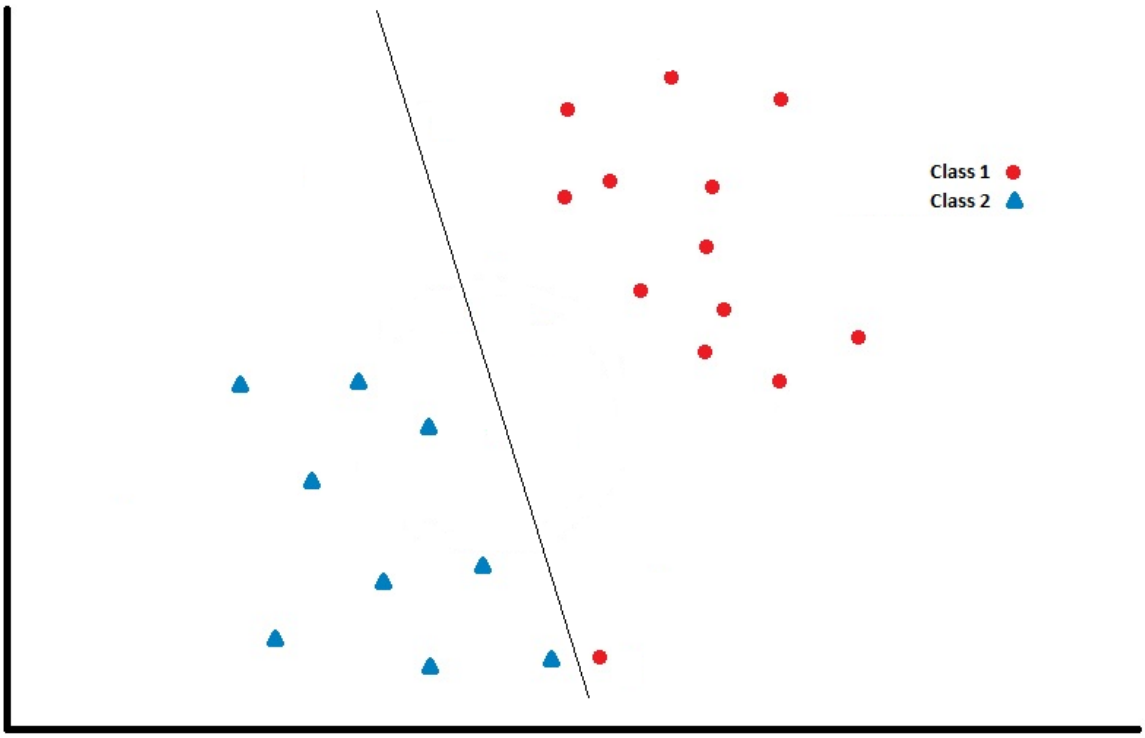

Figure 2. Support Vector Machine Hyperplane

In a real situation it, more often than not isn't possible to separate the examples into classes perfectly and so the algorithm seeks first to determine the best fit decision boundary which maximises accuracy. To overcome this an error tolerance is introduced which allows for the model to tolerate misclassified examples to produce a more suitable margin.

SVM can also produce a nonlinear decision boundary through the use of a method known as Kernel Trick. This can transform existing features into new features which then allows for a curved decision boundary and margin. The Kernel Trick which will be employed in this study is RBF (Radial Basis

Function) which curves the decision boundary by allowing features to have an influence over it, the influence each example has is determined by the position of the decision boundary and the position of that example, essentially weighting each example when making a classification. The weight of each individual training example can also be changed using the gamma parameter.

From the previous studies it can be seen that SVM performs well in a variety of situations, often outperforming other classifiers it is compared against.

SVM is an effective classifier when there is a distinct margin between training data classes which may be the case with student grade data. SVM also performs better with smaller data sets, however large datasets most heavily affect the time taken to train and, while time taken to train should be considered when selecting a model, in this experiment it is not the important, which work well with the data sets available. There are drawbacks to SVM in that performance is generally better with large numbers of features, where the available data sets have less than twenty features for this experiment.

3.4.5 Multi-Layer Perceptron

A Single Layer Perceptron is binary classifier which classifies data by taking in feature data, applying weights to each piece of data and then determining a suitable linear hyperplane to separate the data into appropriate classes.

Multi-Layer Perceptrons are artificial neural networks which consist of a great number of Single Layer Perceptrons acting as neurons. Input data from the features is fed into an initial layer of neurons where they are then modified by nonlinear transfer functions. The outputs of these neurons are weighted and fed into the next layer of neurons where they are further modified by nonlinear transfer functions and passed to the next layer of neurons. Every node in a layer is connected to every node in the next layer with different weights applied to each connection. This process continues until a final output layer is reached. (Gardner & Dorling, 1998)

Multi-Layer perceptrons are extremely flexible and well suited to taking numerical inputs and directly determining outputs regardless of the size of the data set and the number of available feature which is advantageous in this case as data availability will likely differ greatly from course to course. One of the main drawbacks may be the computation time required when training. Another drawback is that, while they often perform well, in the event that they perform poorly it can be difficult to determine the root cause.

3.4.6 Model Evaluation

As mentioned previously, Scikit Learn provides a variety of evaluation methods which can be used to determine the performance of each model. In this experiment students will be classified as either a pass or a fail and the problem to be solved is to identify students at risk of failing. While all evaluation metrics are useful in some way, it is most important to know how often failing students are being misclassified as passing.

The evaluation metrics used rely on knowing how each test example has been classified. This knowledge can be obtained by constructing a confusion matrix for each experiment carried out which will give a 2x2 grid showing whether each training example was classified as true positive, false positive, true negative or false negative where positive is a failing result while negative is a passing result. While confusion matrixes themselves will not be used as metrics, it helps to understand how the values gained from the metrics used are calculated.

One of the most commonly used methods of evaluation is the classification accuracy. This is the number of correctly classified examples divided by the total number of classification attempts using the test data set represented by the following equation:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

This is a good all-purpose evaluation for all classification models which gives a general idea of how classifiers stack up against each other at a glance. While accuracy does give a good idea of general performance, in data sets which are heavily skewed high accuracy may be achieved but the rate of correct predictions from the class with the smaller number of examples may still be extremely low. Accuracy will be used as a performance indicator in this experiment but will be considered a tertiary metric.

Two more relevant evaluation metrics are Precision and Recall.

Precision, also known as the positive predictive value, is the percentage of positive predictions made which are correct and is represented by the following equation:

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

While more relevant than accuracy as an evaluation metric, for this experiment this will be a secondary evaluation metric when considering the rate of false negatives but will remain very useful as a general evaluation metric when considering other factors such as class size as it provides more relevant information when too many false positives can cause problems. In the case of a course with large numbers of students a poor precision score means an educator will spend much more time providing more detailed feedback for those students who do not require it, thus contributing to a lack of timeliness in delivering feedback to those who do need it which has been seen to be detrimental to student performance.

Recall, also known as the true positive rate is the percentage of correctly classified positive examples from the test data set and is represented by the following equation:

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

The recall is arguably the most important evaluation metric available in the context of this problem. The ultimate goal of a machine learning system in predicting student grades is to identify those in need of help. A high rate of false negatives, or a low recall, means that many of the students who are at risk of failing are being identified as passing and, as such will receive unsuitable feedback.

The last of the performance metric which will be considered in this study is the F-1 score, or the harmonic mean of the precision and recall, which is given by the equation:

$$2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The F1-Score is a metric which becomes more useful as course sizes increase and should be taken into account when trying to determine the best classifier as, with smaller student cohorts the greatest cost is incurred when failing students are misclassified as passing, however when student cohorts increase

in size, the cost of misclassifying either passing or failing students begins to even out as available lecturer time has to be taken into account.

Testing with these metrics can be carried out by splitting the available data set into training data, used to train the machine learning model, and test data, used to determine if the model can correctly predict outcomes. How the and where the data is split can result in vastly different evaluation results however, especially in heavily skewed data or smaller data sets.

In order to obtain evaluations which are as accurate as possible it is best to test and train with all data using a variety of different splits. A technique known as K-Fold Cross Validation exists for this exact purpose. The data set is split into a number of sections equal to K, e.g. if K is 10 then the data will be split into 10 equal sections each containing 10% of the data. The first section is selected as the test data set while the rest of the sections are used to train the model, after training the evaluations are carried out. The next section is then selected as test data and the other nine are used as training data. This goes on until all sections have been used to test.

3.5. Feature Importance

Feature selection is an important process in machine learning which can have drastic effects on the overall performance of a machine learning tool. As more features are introduced to a model, it becomes more likely that one or more features will begin to have a detrimental effect on the overall performance of the model. Of course, this is not true for all models so it is not as simple as stating that one feature is causing problems so removing it will improve all models, some complex algorithms can benefit from extra features and data while others require highly correlated data. Determining which features from the available data are most relevant to predicting the outcome and selectively removing the least relevant can increase the model accuracy and reduce overfitting due to irrelevant data and reduce the time spent on training the model due to there being less data to work with.

There were a few methods used to determine the relevance of each feature. Tree based algorithms are well suited to computing feature importance and Scikit Learn provides a function using Random Forests which evaluates the importance of each feature and provides these in an array. The higher the feature importance value assigned, the more important the feature is to the classification.

The feature importances when all available features for the course A and course B student data are taken into consideration are shown below in Table 7 and Table 8.

Course A	
Gender	0.00756783
Region	0.06281934
Highest Education	0.01748003
IMD Band	0.04190574
Age Band	0.01914752
TMA1	0.09935294
TMA2	0.05216397
TMA3	0.18376428
TMA4	0.22085948
TMA5	0.29493889

Table 7: Course A Feature Importance - All Features

Course B	
Gender	0.00777707
Region	0.03715455
Highest Education	0.0120538
IMD Band	0.03533203
Age Band	0.00652249
CMA1	0.01150146
CMA2	0.02948499
CMA3	0.02154071
CMA4	0.06813567
CMA5	0.21310302
TMA1	0.04680448
TMA2	0.04520593
TMA3	0.06327904
TMA4	0.06854425
TMA5	0.21336033
TMA6	0.12020017

Table 8: Course B Feature Importance - All Features

This shows, in almost all cases, that the tutor marked assessment grades are most important in predicting final outcome which is unsurprising as these will required the greatest knowledge in the given course. What is new however, is that student region and IMD band are almost important as the tutor marked assessments and, in some cases, more important than the computer marked assessments. Gender and age band show very low importance which is expected given the previous works showing their lack of correlation with student performance and given they have two and three categories respectively.

It is also advantageous to split the sets of features to see importance without other features. Tables Table 9, Table 10, Table 11 and Table 12 show the importances of student information and grades when separate from each other.

Course A Student Info	
Gender	0.09116148
Region	0.34065359
Highest Education	0.12512761
IMD Band	0.33448967
Age Band	0.10856766

Table 9: Course A Feature Importance - Student Information

Course B Student Info	
Gender	0.04213622
Region	0.5110228
Highest Education	0.09774718
IMD Band	0.28927658
Age Band	0.05981723

Table 10: Course B Feature Importance - Student Information

Course A Grades	
TMA1	0.09184333
TMA2	0.0889804
TMA3	0.11916211
TMA4	0.21408078
TMA5	0.48593339

Table 11: Course A Feature Importance - Grade Information

Course B Grades	
CMA1	0.01607847
CMA2	0.04287737
CMA3	0.02621958
CMA4	0.11867723
CMA5	0.19176719
TMA1	0.0707873
TMA2	0.06974694
TMA3	0.10159528
TMA4	0.07103388
TMA5	0.16154711
TMA6	0.12966963

Table 12: Course B Feature Importance - Grades

Looking at the student information, it can clearly be seen that region and poverty have significantly higher importance than any of the other available features. Gender and age band once again show very low importance while highest education remains firmly in the middle of the others.

The grade information shows a clear increase in feature importance for assessments submitted later in the course. This could be due to an increase in difficulty and relevance making it easier to identify poorer performing students. Interestingly the later computer marked assessments have a higher importance than most of the tutor marked assessments while the tutor marked assessments have a higher overall importance. Again, the higher individual importance could be down to relevance and experience while the tutor marked assessments may be more consistent due to tutor feedback after marking.

Heatmaps were also created using the available data. These show how each feature correlates with each other and with the final outcome label. This allows us to determine how independent each feature is, which is a factor which affects Naïve Bayes models, and will provide more information on the importance of the features to the outcome. Figure 4 shows the heatmap for all available data for course B.

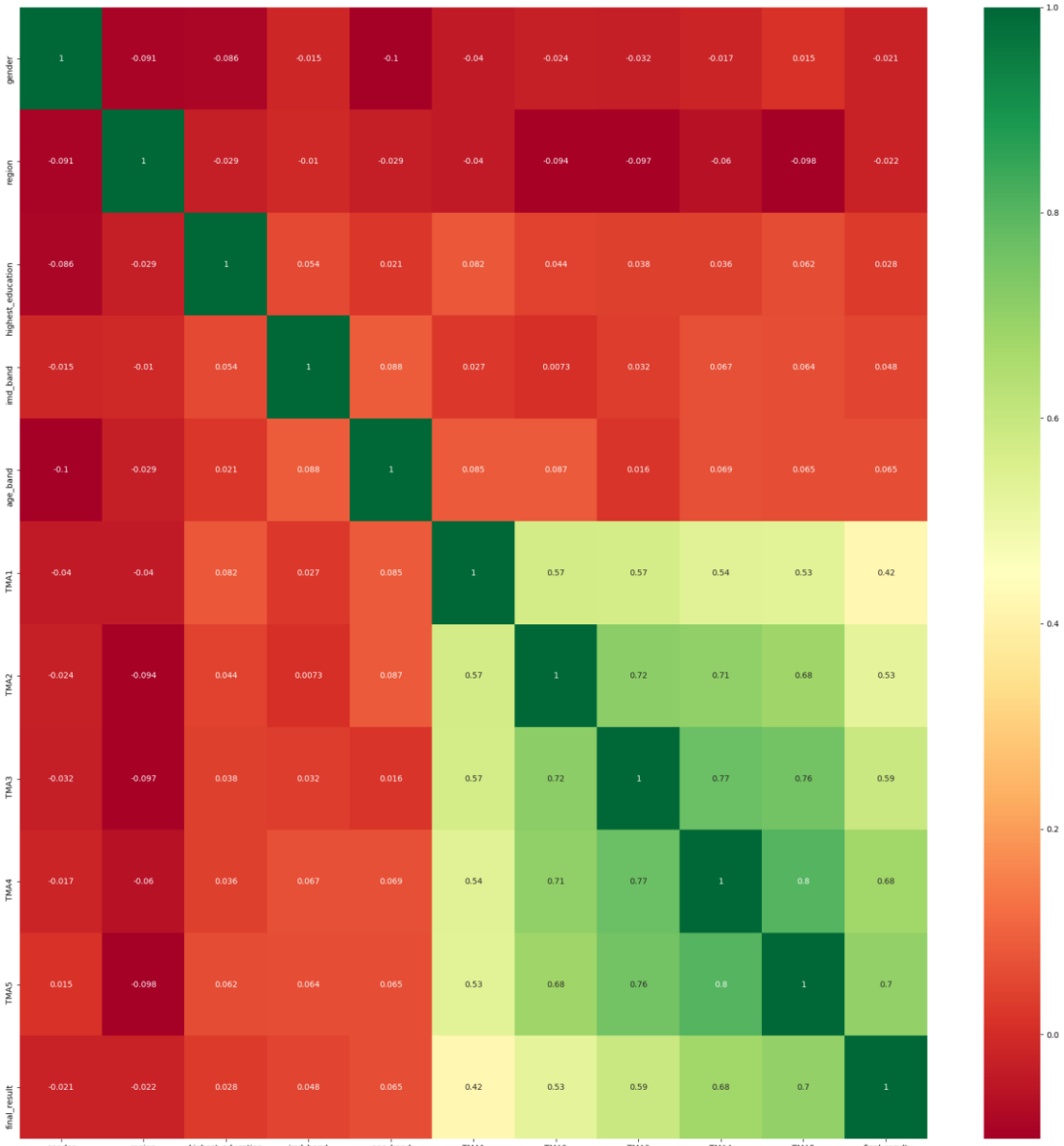

Figure 3: Course A Heatmap

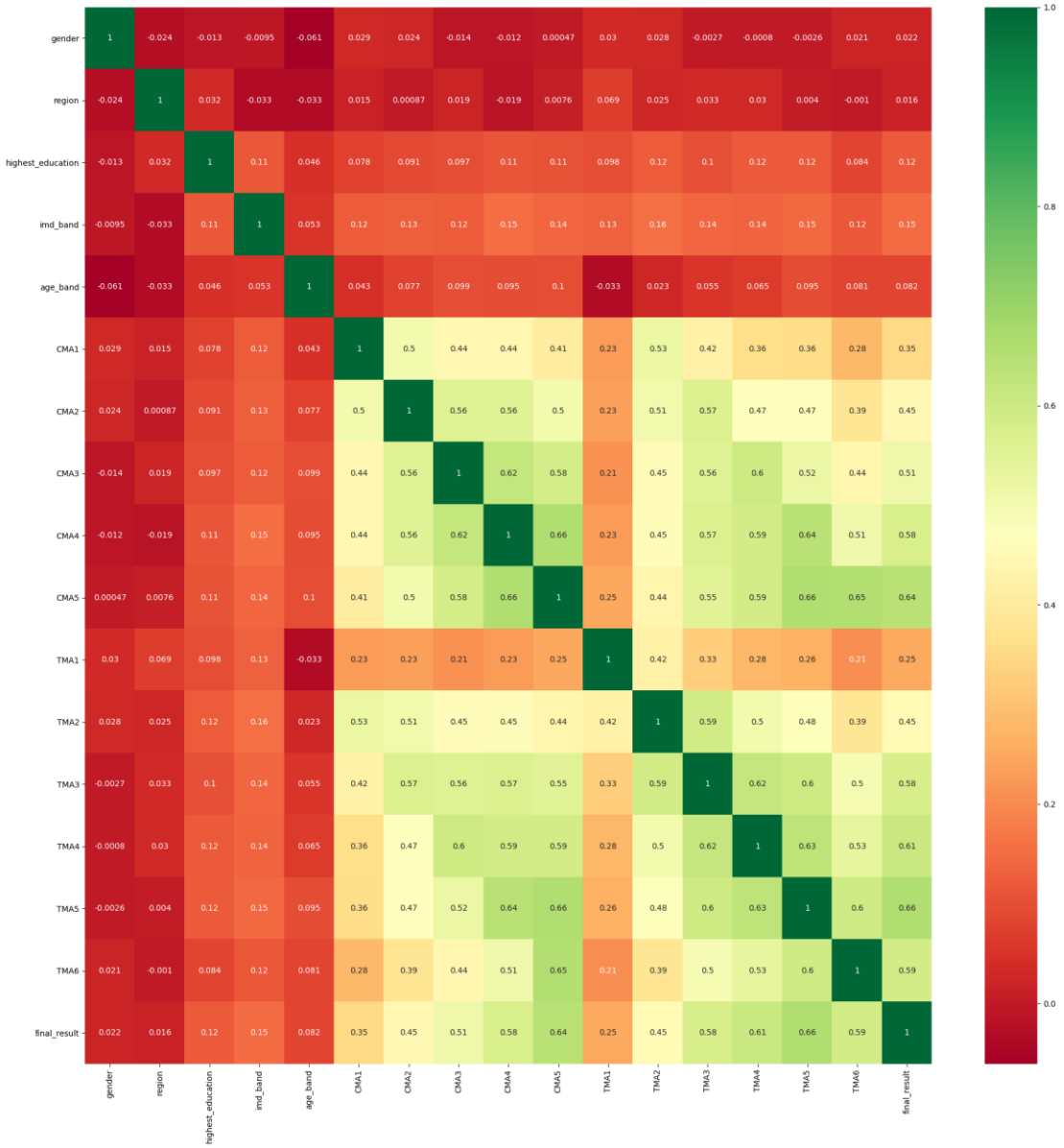

Figure 4: Course B Heatmap

The correlations of each of the features to the final result agree with the feature importances previously shown in that the later assessments correlate more heavily while earlier assessments are still reasonably correlated. The student information is again showing a low correlation with the final result though, where the feature importance showed IMD band and region being as important as some grades, the heat map shows less correlation than the grades. In fact region is least correlated with the final result, with gender remaining very low also. It should be noted that all student information features have a very low correlation to all other features so, despite the low correlation with the final result, retaining these features may benefit algorithms which depend on feature independence.

For Course B TMA1 shows relatively low correlation with the final result and with all other grade features, this may be down to it being early in the course but it may also suggest that the material in the assessment itself is less relevant to the course. Despite this, TMA1 still has a high enough correlation to the final result that it will be considered as necessary, especially when attempting to predict the final result using only earlier grades.

For both courses it can be seen for the rest of the grades that correlation increases with later assessments which supports the feature importances in that later courses are more highly correlated with the final result which is to be expected.

From both methods it can clearly be seen that the feature most likely to reduce performance is a student's gender as it holds very low importance and correlation in all situations. It may also be possible for model performance to improve with the removal of age band which similarly has low importance and correlation or with the removal of student region as it has an extremely low correlation. Tests will be carried out with these removed individually to determine if there is any significant change in performance.

4. Analysis

This chapter will describe and discuss the results of the experiment after having processed the data into a usable state and having implemented the application.

4.1. Results

As mentioned, the classifiers being compared are Gaussian Naïve Bayes, K-Nearest Neighbours, Random Forest, SVM and Multi-Layer Perceptron. The data sets used are Course A and Course B. The results of each set of experiments will be provided in tables as mean scores of the results gained through 10-fold cross validation.

Each model was first tested using the full dataset, the student information only and then the grades only. Next the models were tested with the removal each of the three student information features identified as possibly negatively affecting the performance of models having been removed individually and then with all three removed. Tests were then carried out with all student information and one, two, three and four grades, or pairs of grades in the case of Course B.

4.1.1 Gaussian Naïve Bayes

As Gaussian Naïve Bayes possesses no tuning parameters only one set of experiments using each data set was carried out.

With the Course A data set Gaussian Naïve Bayes performs poorly as a classifier. Accuracy is high however, this is likely very much due to the data being skewed in favour of passes. Any single false negative results in a large drop in recall but a near negligible drop in accuracy. Removing features which were deemed unimportant results in almost no drop in accuracy, in fact using only student data results in a complete failure to accurately identify any failing examples. As grades are added as features, the precision and F1-Score steadily rise, however the recall remains relatively unchanged after using only two early grades. Grade data can be seen as the only important data in this data set when using this algorithm.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.911546339	0.754300144	0.679898157	0.702633994
Student Information Only	0.853507	0	0	0
Grades Only	0.911546339	0.754300144	0.679898157	0.702633994
Gender Removed	0.913133641	0.754300144	0.686564824	0.706555563
Age Removed	0.911546339	0.754300144	0.679898157	0.702633994
Region Removed	0.911546339	0.754300144	0.679898157	0.702633994
Age, Gender and Region Removed	0.911546339	0.754300144	0.679898157	0.702633994
TMA2 – TMA5 Removed	0.868049155	0.23479798	0.518333333	0.313738418
TMA3 - TMA5 Removed	0.884178187	0.464574315	0.681746032	0.529577498
TMA4 - TMA5 Removed	0.889042499	0.582655123	0.648683261	0.584936551
TMA5 Removed	0.903558628	0.686479076	0.663603896	0.658116883

Table 13: Evaluation Results for Gaussian Naive Bayes Experiments with Course A Data

The Course B results fare much better, with significantly higher recall rates. Removal of unimportant features results in a negligible change in all evaluation metrics. Student information alone gives poor precision but a surprisingly high recall score. Again, using only one or two early grades brings the recall

to a similar score as with a higher number of grades, however we do see the precision steadily increase.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.879545	0.771786	0.815241	0.79195
Student Information Only	0.657683	0.239767	0.507051	0.324418
Grades Only	0.877659	0.770399	0.812008	0.789721
Gender Removed	0.879813	0.772655	0.815367	0.792538
Age Removed	0.878466365	0.771327134	0.813031622	0.790603639
Region Removed	0.878736632	0.771227706	0.813734616	0.790820862
Age, Gender and Region Removed	0.87846709	0.770256832	0.813518982	0.790236578
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.768524476	0.363789609	0.727397244	0.484063824
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.797090108	0.539277348	0.711952327	0.612432776
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.820800365	0.656135772	0.719343797	0.684650438
TMA5 – TMA6 and CMA5 Removed	0.839390778	0.712455728	0.742204374	0.72509174

Table 14: Evaluation Results for Gaussian Naive Bayes Experiments with Course B Data

From these results it seems that a larger dataset had the most effect on the Recall score for Gaussian Naïve Bayes as the Course B results show significantly higher recall scores in all situations. Reducing the number of features has marginal effect in all but the most extreme situation where grades were not taken into account, and even in this scenario Course A misclassified all failing students while Course B resulted in only half of all failing students being misclassified.

Precision was similar for both data sets with higher numbers of features and was dramatically affected by reducing the number of important features. Removing features of lower importance had a negligible effect on any of the performance metrics.

4.1.2 K-Nearest Neighbours

K-Nearest neighbours possesses the parameter K with which the number of nearest neighbours can be determined in order to tune the model. Experiments were carried out with K equal to 1, 2, 5, 10 and 20 to determine if and how this varied the performance.

From Table 15 and Table 16, it can be seen that, while using the Course A results, the performance saw a slight dip at the lowest K values. Both Precision and Recall were severely affected by the reduction in important features, with the removal of more important features such as region and grades.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.909831029	0.657463925	0.724254912	0.679850488
Student Information Only	0.747644649	0.155454545	0.126126374	0.12808866
Grades Only	0.898566308	0.694451659	0.671462704	0.677648655
Gender Removed	0.909831029	0.657463925	0.724254912	0.679850488
Age Removed	0.909831029	0.657463925	0.724254912	0.679850488
Region Removed	0.89859191	0.680165945	0.648313353	0.655124224
Age, Gender and Region Removed	0.89859191	0.680165945	0.648313353	0.655124224
TMA2 – TMA5 Removed	0.776548899	0.265353535	0.233225108	0.229829936
TMA3 - TMA5 Removed	0.826369688	0.470046898	0.426282051	0.423416149
TMA4 - TMA5 Removed	0.843881208	0.464220779	0.477056277	0.455517483
TMA5 Removed	0.882565284	0.606580087	0.58998557	0.581051269

Table 15: Evaluation Results for K-Nearest Neighbours Experiments with Course A Data and K = 1

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.879288274	0.775808081	0.561232232	0.644907438
Student Information Only	0.65266257	0.311937229	0.144745843	0.188595166
Grades Only	0.860035842	0.778585859	0.511965812	0.606932851
Gender Removed	0.879288274	0.775808081	0.561232232	0.644907438
Age Removed	0.880901178	0.775808081	0.566360437	0.647805988
Region Removed	0.869738863	0.778585859	0.535374674	0.623605099
Age, Gender and Region Removed	0.869738863	0.778585859	0.5360157	0.623736851
TMA2 – TMA5 Removed	0.723604711	0.335050505	0.209124209	0.24371211
TMA3 - TMA5 Removed	0.795903738	0.603827561	0.389592652	0.455045767
TMA4 - TMA5 Removed	0.813364055	0.568279221	0.409334979	0.46371996
TMA5 Removed	0.855248336	0.722027417	0.504902646	0.579848791

Table 16: Evaluation Results for K-Nearest Neighbours Experiments with Course A Data and K = 2

Tables Table 17, Table 18 and Table 19 show results using higher K values. There is a clear increase in Recall scores as K values are increased, with a possible saturation after K=10 however, the precision score sees a significantly larger drop than the gain in Recall resulting in an overall drop in F1-Score.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.922708653	0.60011544	0.87265873	0.699491678
Student Information Only	0.835867896	0.02020202	0.1	0.033566434
Grades Only	0.929160266	0.640916306	0.874444444	0.727530364
Gender Removed	0.922708653	0.60011544	0.87265873	0.699491678
Age Removed	0.922708653	0.60011544	0.87265873	0.699491678
Region Removed	0.929160266	0.640916306	0.874444444	0.727530364
Age, Gender and Region Removed	0.929160266	0.640916306	0.874444444	0.727530364
TMA2 – TMA5 Removed	0.84718382	0.205353535	0.47452381	0.268506787
TMA3 - TMA5 Removed	0.888914491	0.412806638	0.7675	0.488633763
TMA4 - TMA5 Removed	0.900204813	0.444372294	0.807857143	0.554095202
TMA5 Removed	0.91953405	0.568802309	0.843492063	0.671092059

Table 17: Evaluation Results for K-Nearest Neighbours Experiments with Course A Data and K = 5

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.92593446	0.573600289	0.938928571	0.696761134
Student Information Only	0.848694316	0.009090909	0.05	0.015384615
Grades Only	0.92109575	0.573600289	0.894920635	0.68475975
Gender Removed	0.92593446	0.573600289	0.938928571	0.696761134
Age Removed	0.92593446	0.573600289	0.938928571	0.696761134
Region Removed	0.922708653	0.573600289	0.909206349	0.688681319
Age, Gender and Region Removed	0.922708653	0.573600289	0.909206349	0.688681319
TMA2 – TMA5 Removed	0.863184844	0.216464646	0.502857143	0.29121741
TMA3 - TMA5 Removed	0.8937532	0.426948052	0.766190476	0.528075666
TMA4 - TMA5 Removed	0.897004608	0.42492785	0.821428571	0.527609407
TMA5 Removed	0.921121352	0.549357864	0.870833333	0.662221164

Table 18: Evaluation Results for K-Nearest Neighbours Experiments with Course A Data and K = 10

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.924347158	0.52257215	0.953214286	0.664212571
Student Information Only	0.853507424	0	0	0
Grades Only	0.924347158	0.52257215	0.953214286	0.664212571
Gender Removed	0.924347158	0.52257215	0.953214286	0.664212571
Age Removed	0.924347158	0.52257215	0.953214286	0.664212571
Region Removed	0.924347158	0.52257215	0.953214286	0.664212571
Age, Gender and Region Removed	0.924347158	0.52257215	0.953214286	0.664212571
TMA2 – TMA5 Removed	0.869636457	0.169747475	0.633333333	0.251648352
TMA3 - TMA5 Removed	0.892140297	0.37023088	0.828333333	0.47958277
TMA4 - TMA5 Removed	0.901792115	0.395761183	0.925	0.526502883
TMA5 Removed	0.917921147	0.527766955	0.870833333	0.642910409

Table 19: Evaluation Results for K-Nearest Neighbours Experiments with Course A Data and K = 20

Table 20 and Table 21 again show low K values with the Course B data set. There is only a marginal improvement in Recall score with the larger data set, with a similar reduction moving from K=1 to K=2. Precision shows a larger improvement than with the Course A data set but not significantly so.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.845333015	0.736745478	0.743334679	0.739030608
Student Information Only	0.575566751	0.375769026	0.381720791	0.377665243
Grades Only	0.847487175	0.738328713	0.7477307	0.74209466
Gender Removed	0.845333015	0.736745478	0.743334679	0.739030608
Age Removed	0.845333015	0.736745478	0.743334679	0.739030608
Region Removed	0.847487175	0.737107134	0.748845867	0.741813544
Age, Gender and Region Removed	0.847756717	0.737926806	0.748999929	0.742307354
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.654537287	0.430566473	0.422839038	0.424845962
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.710859925	0.524780274	0.514336409	0.518518912
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.767988291	0.59966128	0.614567179	0.605124598
TMA5 – TMA6 and CMA5 Removed	0.800316639	0.663283352	0.66592619	0.663047113

Table 20: Evaluation Results for K-Nearest Neighbours Experiments with Course B Data and K = 1

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.80895212	0.791885926	0.646661735	0.710442021
Student Information Only	0.52115869	0.587553554	0.374955999	0.456602667
Grades Only	0.809489755	0.795224064	0.647233454	0.711911912
Gender Removed	0.809220937	0.791885926	0.647106246	0.710712571
Age Removed	0.809221662	0.791885926	0.647143238	0.710742817
Region Removed	0.809760021	0.792308715	0.648270059	0.711522499
Age, Gender and Region Removed	0.810029563	0.792308715	0.648736997	0.711793266
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.593635336	0.618740017	0.387538677	0.474943299
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.654804655	0.657398403	0.447395957	0.530888497
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.721911138	0.728587352	0.524267417	0.608515652
TMA5 – TMA6 and CMA5 Removed	0.775257949	0.770613825	0.594660588	0.669885406

Table 21: Evaluation Results for K-Nearest Neighbours Experiments with Course B Data and K = 2

Tables Table 22, Table 23 and Table 24 show a significant jump in Recall score at K=5 with only slight increases thereafter. The Precision score remains unchanged across higher K values which allows for marginal F1 score improvements.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.880089413	0.714014336	0.861412559	0.779653267
Student Information Only	0.622166247	0.31557923	0.432673073	0.363712506
Grades Only	0.879819871	0.714789529	0.85982232	0.7794986
Gender Removed	0.880089413	0.714014336	0.861412559	0.779653267
Age Removed	0.880089413	0.714014336	0.861412559	0.779653267
Region Removed	0.879280787	0.713159635	0.859460308	0.778243419
Age, Gender and Region Removed	0.879550329	0.713159635	0.860464324	0.778692828
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.726483929	0.403184828	0.558703421	0.466894773
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.772030693	0.487905095	0.661107557	0.560024607
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.814876243	0.57496373	0.749646693	0.648731866
TMA5 – TMA6 and CMA5 Removed	0.851522331	0.642605925	0.823453811	0.720690692

Table 22: Evaluation Results for K-Nearest Neighbours Experiments with Course B Data and K = 5

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.883860099	0.72759808	0.863109745	0.787808946
Student Information Only	0.618387909	0.38267232	0.438883196	0.407627121
Grades Only	0.884130365	0.727475243	0.864275994	0.788486558
Gender Removed	0.884129641	0.728411088	0.863232926	0.788375471
Age Removed	0.883860099	0.72759808	0.863109745	0.787808946
Region Removed	0.884669449	0.729224096	0.864462172	0.789517783
Age, Gender and Region Removed	0.884399907	0.728411088	0.864341558	0.788956432
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.732679042	0.42391103	0.571399139	0.485617416
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.777157059	0.512882098	0.66430996	0.577639098
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.827273715	0.604092726	0.77183744	0.675877991
TMA5 – TMA6 and CMA5 Removed	0.852328058	0.649611885	0.821840638	0.723855966

Table 23: Evaluation Results for K-Nearest Neighbours Experiments with Course B Data and K = 10

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.890058111	0.716814695	0.896164221	0.79490974
Student Information Only	0.644584383	0.313110977	0.472579217	0.375568366
Grades Only	0.889519752	0.71594513	0.895053093	0.793954316
Gender Removed	0.890058111	0.716814695	0.896164221	0.79490974
Age Removed	0.890058111	0.716814695	0.896164221	0.79490974
Region Removed	0.890058111	0.716814695	0.896164221	0.79490974
Age, Gender and Region Removed	0.890058111	0.716814695	0.896164221	0.79490974
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.756664638	0.390365997	0.656983832	0.488039114
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.795477205	0.486840417	0.741398039	0.586056798
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.834009362	0.578775794	0.815312776	0.67403841
TMA5 – TMA6 and CMA5 Removed	0.858257978	0.633787255	0.85732573	0.726852583

Table 24: Evaluation Results for K-Nearest Neighbours Experiments with Course B Data and K = 20

K-Nearest Neighbour's Recall score seems to benefit more from a smaller data set and a smaller number of features in the experiments above with almost all Course A experiments outperforming those of Course B. Precision score suffers quite heavily with the loss of important features, especially

in situations where the Recall score benefits i.e. with smaller datasets and high values of K. The F1 score seems to benefit from the larger data set as the Precision and Recall are much more balanced.

4.1.3 Random Forest

Tuning of Random Forest classifiers is dependent on the number of trees in the model *n*. In this case experiments were carried out with *n* equal to 4, 8, 16 and 32 to determine how increasing the number of trees affects performance.

Tables Table 25, Table 26, Table 27 and Table 28 hold the results for experiments carried out using the Course A data set. The Recall score sees a clear increase with diminishing returns as the number of trees is increased, with very little improvement between n=16 and n=32. The Precision score sees a small decrease as the number of trees increases. With higher *n* values the Recall score is much more tolerant to losing important features, Precision has no such tolerance and is affected heavily in all cases.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.905069124	0.720487013	0.679935065	0.67269178
Student Information Only	0.729953917	0.140725108	0.114722222	0.116361474
Grades Only	0.909856631	0.734098124	0.677727273	0.695116622
Gender Removed	0.919559652	0.753189033	0.718517316	0.722767684
Age Removed	0.8937532	0.691908369	0.632647908	0.650354582
Region Removed	0.901792115	0.664047619	0.695367965	0.66763147
Age, Gender and Region Removed	0.8906298	0.691800144	0.59965812	0.636117588
TMA2 – TMA5 Removed	0.810368664	0.351753247	0.382446304	0.33685142
TMA3 - TMA5 Removed	0.835995904	0.440324675	0.457713675	0.42881717
TMA4 - TMA5 Removed	0.847388633	0.592857143	0.507735043	0.513955466
TMA5 Removed	0.885893497	0.667943723	0.604519925	0.613594004

Table 25: Evaluation Results for Random Forest Experiments with Course A Data and n = 4

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.913082437	0.641637807	0.74284188	0.677597671
Student Information Only	0.760419867	0.096237374	0.106581197	0.08725957
Grades Only	0.916282642	0.647741703	0.760984848	0.687988925
Gender Removed	0.917946749	0.683340548	0.763322511	0.70603393
Age Removed	0.92749616	0.709603175	0.824603175	0.760068762
Region Removed	0.92437276	0.711244589	0.780912698	0.730396724
Age, Gender and Region Removed	0.921121352	0.654047619	0.808441558	0.707949945
TMA2 – TMA5 Removed	0.837608807	0.30713925	0.433095238	0.331540795
TMA3 - TMA5 Removed	0.84889913	0.428751804	0.491515152	0.43070114
TMA4 - TMA5 Removed	0.887429595	0.521284271	0.691587302	0.562997679
TMA5 Removed	0.898694316	0.633726551	0.686388889	0.631745052

Table 26: Evaluation Results for Random Forest Experiments with Course A Data and n = 8

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.930747568	0.693423521	0.843008658	0.757049274
Student Information Only	0.79078341	0.098023088	0.22202381	0.108957688
Grades Only	0.929083461	0.691673882	0.843481241	0.755173547
Gender Removed	0.930773169	0.690645743	0.847175325	0.754794372
Age Removed	0.933998976	0.695569986	0.853088023	0.759599617
Region Removed	0.933973374	0.690645743	0.865595238	0.760290463
Age, Gender and Region Removed	0.930798771	0.667034632	0.843567821	0.733035788
TMA2 – TMA5 Removed	0.847132616	0.230775613	0.463333333	0.282753358
TMA3 - TMA5 Removed	0.874577573	0.436580087	0.642619048	0.488223679
TMA4 - TMA5 Removed	0.887378392	0.495887446	0.71010101	0.548700412
TMA5 Removed	0.914695341	0.628405483	0.790025253	0.676060606

Table 27: Evaluation Results for Random Forest Experiments with Course A Data and n = 16

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.930773169	0.655526696	0.867691198	0.738877665
Student Information Only	0.794086022	0.086911977	0.132619048	0.093717949
Grades Only	0.937224782	0.691673882	0.883531746	0.770142409
Gender Removed	0.932360471	0.685487013	0.855786436	0.754657718
Age Removed	0.925908858	0.665284993	0.827849928	0.729850026
Region Removed	0.932386073	0.669054834	0.872420635	0.751293363
Age, Gender and Region Removed	0.934024578	0.692431457	0.840551948	0.750224432
TMA2 – TMA5 Removed	0.853584229	0.234437229	0.478333333	0.296178266
TMA3 - TMA5 Removed	0.885765489	0.419119769	0.723452381	0.49469973
TMA4 - TMA5 Removed	0.885791091	0.47012987	0.715952381	0.525369532
TMA5 Removed	0.914720942	0.601378066	0.801111111	0.66401475

Table 28: Evaluation Results for Random Forest Experiments with Course A Data and n = 32

Tables Table 29, Table 30, Table 31 and Table 32 hold the results for experiments carried out using the Course B data set. Performance trends are very similar to those seen with Course A data. The Recall score

The Recall score has a clear increase with diminishing returns as the number of trees is increased, with very little improvement between n=16 and n=32. The Precision score sees a small decrease as the number of trees increases, though the decrease becomes negligible at higher values for *n* and given the F-1 score increases, it can be taken that the trade-off is not too severe.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.855835724	0.782022955	0.746738323	0.762586153
Student Information Only	0.613853904	0.359316224	0.42545775	0.388592964
Grades Only	0.845059125	0.770119792	0.727965588	0.747230336
Gender Removed	0.852599049	0.76374576	0.748794093	0.754884669
Age Removed	0.851793322	0.782099792	0.737962762	0.758519856
Region Removed	0.853946034	0.77008155	0.747881315	0.75703088
Age, Gender and Region Removed	0.849090659	0.785558033	0.729913396	0.75540221
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.672057502	0.500142135	0.454903734	0.474182881
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.718141901	0.59514514	0.523916308	0.556176845
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.780921224	0.675156959	0.623256763	0.647034461
TMA5 – TMA6 and CMA5 Removed	0.803013506	0.718578039	0.655099051	0.684572831

Table 29: Evaluation Results for Random Forest Experiments with Course B Data and n = 4

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.875236936	0.754326517	0.818478198	0.783571235
Student Information Only	0.604282116	0.346341244	0.411420473	0.374314984
Grades Only	0.872807437	0.747917553	0.813237067	0.777543236
Gender Removed	0.877659189	0.754183941	0.823948703	0.785117641
Age Removed	0.876846216	0.758833262	0.818047369	0.786432599
Region Removed	0.874965945	0.758956225	0.81150641	0.783358008
Age, Gender and Region Removed	0.88197548	0.758852043	0.832641509	0.792604347
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.696309017	0.469417911	0.490869524	0.478567842
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.744537432	0.540394356	0.579297874	0.557517443
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.806791438	0.635455894	0.696631108	0.661791093
TMA5 – TMA6 and CMA5 Removed	0.841552908	0.68346779	0.761921106	0.719127442

Table 30: Evaluation Results for Random Forest Experiments with Course B Data and n = 8

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.887361969	0.737869633	0.866070894	0.795881982
Student Information Only	0.608312343	0.330878676	0.412299027	0.366065733
Grades Only	0.885749065	0.736490807	0.86222476	0.792989236
Gender Removed	0.88735907	0.737711504	0.866221925	0.795701268
Age Removed	0.88951468	0.746723821	0.8660529	0.800568181
Region Removed	0.889248036	0.737515395	0.873774903	0.798535104
Age, Gender and Region Removed	0.883857201	0.736434336	0.856235199	0.790393368
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.712209808	0.439857108	0.520722709	0.475603244
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.772834971	0.529428047	0.646826527	0.580625679
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.830235052	0.638346603	0.758830743	0.691759444
TMA5 – TMA6 and CMA5 Removed	0.851249891	0.662215971	0.810350485	0.727180193

Table 31: Evaluation Results for Random Forest Experiments with Course B Data and n = 16

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.88844086	0.730799451	0.877358283	0.796307189
Student Information Only	0.610327456	0.330220711	0.41554297	0.366747851
Grades Only	0.890325479	0.732757104	0.880149581	0.798404436
Gender Removed	0.891403646	0.727444988	0.890869626	0.799808482
Age Removed	0.887360519	0.731536438	0.871438546	0.794037747
Region Removed	0.892211547	0.733277567	0.887152442	0.801615481
Age, Gender and Region Removed	0.888711851	0.732978699	0.876107306	0.796969515
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.717056488	0.424040406	0.535212749	0.471528205
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.780384314	0.508169603	0.677194964	0.579824873
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.831044402	0.606387171	0.779902402	0.679906713
TMA5 – TMA6 and CMA5 Removed	0.857450801	0.653721463	0.838721906	0.732821308

Table 32: Evaluation Results for Random Forest Experiments with Course B Data and n = 32

Random Forest performs well with either small or large data sets as can be seen from the results above with marginal performance increases with larger data sets when there are enough important features available. With the removal of important features, the Precision scores for Course A dropped significantly more than those for Course B suggesting that the magnitude of this effect may be dependent on the data set size. Due to the increase in number of trees negatively affecting Precision score to a greater degree in larger data sets, and assuming a larger data set is due to greater numbers of students in a cohort, for the purposes of identifying failing students it may actually be beneficial to restrict the number of trees as the cost of false positives increases in larger courses.

The performance changes from the removal of less important features, whether individually or in groups, results in small, almost negligible, changes in all performance metrics. For the most part, when the feature removed has very low performance, this change is positive however, several situations result in a performance drop. This change is small and unpredictable and should only really be considered if it is necessary to increase training times.

4.1.4 Support Vector Machine

Tuning of SVM classifiers is most often carried out in Scikit Learn by altering the value of the cost of misclassified examples , *C*, and the gamma of the kernel. After some investigation it was found that tuning using *C* produced negligible change for these data sets, while tuning gamma resulted in significant performance differences. As such the experiments were conducted using the default *C* value and gamma values of 0.1, 0.01 and 0.001.

Tables Table 33, Table 34 and Table 35 hold the results for experiments carried out using the Course A data set. Recall scores see a sharp increase with the first reduction to the gamma with a minor increase to the Precision score. With the second gamma reduction there is a very large jump in Precision score with a relatively small, but still significant, drop in Recall. Precision suffers heavily when important features are removed while Recall seems to remain relatively high however, this holds true

until too many important features are removed, at which point all positive examples are misclassified. This may be due to the heavily skewed nature of this data set. Removal of less relevant student information has only negligible effect on each of the performance metrics.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.863159242	0.095	0.5	0.155081585
Student Information Only	0.853507424	0	0	0
Grades Only	0.874423963	0.168502886	0.85	0.27528083
Gender Removed	0.863159242	0.095	0.5	0.155081585
Age Removed	0.863159242	0.095	0.5	0.155081585
Region Removed	0.867997952	0.124292929	0.65	0.200769231
Age, Gender and Region Removed	0.867997952	0.124292929	0.65	0.200769231
TMA2 – TMA5 Removed	0.861571941	0.111161616	0.55	0.171208791
TMA3 - TMA5 Removed	0.858320533	0.095	0.45	0.153982684
TMA4 - TMA5 Removed	0.863159242	0.095	0.5	0.155081585
TMA5 Removed	0.863159242	0.095	0.5	0.155081585

Table 33: Evaluation Results for SVM Experiments with Course A Data and gamma = 0.1

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.880824373	0.225970418	0.933333333	0.354997225
Student Information Only	0.853507424	0	0	0
Grades Only	0.890501792	0.316020924	0.925	0.452905983
Gender Removed	0.880824373	0.225970418	0.933333333	0.354997225
Age Removed	0.880824373	0.225970418	0.933333333	0.354997225
Region Removed	0.884024578	0.279505772	0.9	0.408440448
Age, Gender and Region Removed	0.884024578	0.279505772	0.9	0.408440448
TMA2 – TMA5 Removed	0.86641065	0.149545455	0.625	0.224725275
TMA3 - TMA5 Removed	0.893778802	0.35523088	0.841666667	0.481412656
TMA4 - TMA5 Removed	0.884050179	0.272756133	0.91	0.397545788
TMA5 Removed	0.885637481	0.296966089	0.908333333	0.423536464

Table 34: Evaluation Results for SVM Experiments with Course A Data and gamma = 0.01

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.922708653	0.600873016	0.870833333	0.688222026
Student Information Only	0.853507	0	0	0
Grades Only	0.922708653	0.600873016	0.870833333	0.688222026
Gender Removed	0.922708653	0.600873016	0.870833333	0.688222026
Age Removed	0.922708653	0.600873016	0.870833333	0.688222026
Region Removed	0.922708653	0.600873016	0.870833333	0.688222026
Age, Gender and Region Removed	0.922708653	0.600873016	0.870833333	0.688222026
TMA2 – TMA5 Removed	0.871223758	0.158636364	0.7	0.251684982
TMA3 - TMA5 Removed	0.892140297	0.378564214	0.816428571	0.485138326
TMA4 - TMA5 Removed	0.901792115	0.42492785	0.871428571	0.540401191
TMA5 Removed	0.917921147	0.531176046	0.882380952	0.644110276

Table 35: Evaluation Results for SVM Experiments with Course A Data and gamma = 0.001

Tables Table 36, Table 37 and Table 38 hold the results for experiments carried out using the Course B data set. For these experiments the Recall scores were highest, in some cases perfectly classifying all failing students correctly, with a higher gamma value and dropped when it was lowered. Precision scores react similarly to Course A in that they rise as the gamma is lowered with the rate of increase in the score being almost the same. Again, the drop in performance from the removal of more important features is noticeable but not too severe until there are not enough important features, at which point performance drops off severely. Removal has negligible effect on any of the performance metrics for Course B.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.741853607	0.136588745	1	0.23974558
Student Information Only	0.661712846	0.109926958	0.56935814	0.180855948
Grades Only	0.769344695	0.230168379	0.994117647	0.372606833
Gender Removed	0.741853607	0.136588745	1	0.23974558
Age Removed	0.742662232	0.139434727	1	0.244151464
Region Removed	0.756137872	0.185470389	0.99375	0.31132924
Age, Gender and Region Removed	0.756676956	0.187400359	0.99375	0.313900373
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.738616207	0.312224891	0.625368962	0.414231254
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.741583341	0.192831614	0.769481889	0.307315773
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.743201316	0.141213495	1	0.246807997
TMA5 – TMA6 and CMA5 Removed	0.741853607	0.136588745	1	0.23974558

Table 36: Evaluation Results for SVM Experiments with Course B Data and gamma = 0.1

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.793597658	0.313351872	0.989467593	0.474876092
Student Information Only	0.655416	0	0	0
Grades Only	0.799525404	0.336432642	0.985945517	0.500045362
Gender Removed	0.793597658	0.313351872	0.989467593	0.474876092
Age Removed	0.7938672	0.314322745	0.98956229	0.47600899
Region Removed	0.798179144	0.330747372	0.987742053	0.49394883
Age, Gender and Region Removed	0.798179144	0.330747372	0.987742053	0.49394883
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.759632496	0.348810026	0.696603521	0.463125297
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.782813813	0.465981115	0.712573619	0.560997125
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.794133843	0.387659477	0.838095792	0.527246247
TMA5 – TMA6 and CMA5 Removed	0.797369794	0.334083181	0.970749747	0.495585244

Table 37: Evaluation Results for SVM Experiments with Course B Data and gamma = 0.01

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.869584529	0.6200491	0.923886736	0.740214135
Student Information Only	0.655416	0	0	0
Grades Only	0.870662696	0.623568968	0.924321984	0.742970863
Gender Removed	0.869584529	0.6200491	0.923886736	0.740214135
Age Removed	0.869584529	0.6200491	0.923886736	0.740214135
Region Removed	0.870393154	0.623533939	0.92319448	0.74252273
Age, Gender and Region Removed	0.870393154	0.623533939	0.92319448	0.74252273
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.761789555	0.309306073	0.743946564	0.435340136
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.800057966	0.483142771	0.761602419	0.58937054
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.833198562	0.572453517	0.818108997	0.671235538
TMA5 – TMA6 and CMA5 Removed	0.857718894	0.650112309	0.840646386	0.731815163

Table 38: Evaluation Results for SVM Experiments with Course B Data and gamma = 0.001

SVM seems to have a better overall performance with a larger data set and more heavily correlated features. The performance in both data sets remains relatively good even with a very small number of features highly correlated with the outcome. Using student information benefits neither situation in this case and, for the sake of simplification, would be best left out when using SVM. From the experiments conducted it is clear that the effects of changing the gamma are significantly different on different data sets. More experiments would need to be conducted to if there is a common trend as the parameters are tuned.

4.1.5 Multi-Layer Perceptron

In this experiment the tuning of the Multi-Layer Perceptron was done by altering the size of the hidden layer. The experiments were conducted using the hidden layer sizes of 100, 250 and 500.

Tables Table 39, Table 40 and Table 41 hold the results for experiments carried out using the Course A data set. The Recall score is almost always best when using more available features while the precision score is severely negatively affected by the removal of important features. Increasing the hidden layer size seemed to have very little effect on any of the performance metrics, there were slight positive and negative variations but no concrete trends. Training times were extremely slow compared with other algorithms, especially with a larger hidden layer size, often upwards of ten seconds.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.925908858	0.571580087	0.902380952	0.683082707
Student Information Only	0.853507	0	0	0
Grades Only	0.922708653	0.602893218	0.865277778	0.691128009
Gender Removed	0.922734255	0.629408369	0.831666667	0.692985348
Age Removed	0.92109575	0.600386003	0.869325397	0.67578954
Region Removed	0.92109575	0.593044733	0.8675	0.680717178
Age, Gender and Region Removed	0.927547363	0.605310245	0.905992063	0.701177208
TMA2 – TMA5 Removed	0.86641065	0.158636364	0.623333333	0.238113553
TMA3 - TMA5 Removed	0.879237071	0.332604618	0.765833333	0.420716827
TMA4 - TMA5 Removed	0.900256016	0.433665224	0.835833333	0.52854817
TMA5 Removed	0.908269329	0.531807359	0.825	0.607966573

Table 39: Evaluation Results for MLP Experiments with Course A Data and hidden layer size = 100

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.924347158	0.61261544	0.860992063	0.690035572
Student Information Only	0.853507	0	0	0
Grades Only	0.919508449	0.558448773	0.856944444	0.660541928
Gender Removed	0.921070148	0.623095238	0.826904762	0.685951162
Age Removed	0.92109575	0.62511544	0.829325397	0.684602116
Region Removed	0.927521761	0.638012266	0.865873016	0.717846766
Age, Gender and Region Removed	0.92109575	0.624484127	0.831547619	0.695759194
TMA2 – TMA5 Removed	0.863184844	0.158636364	0.556666667	0.233731269
TMA3 - TMA5 Removed	0.882514081	0.346619769	0.751428571	0.44022311
TMA4 - TMA5 Removed	0.900204813	0.44262987	0.835	0.547986523
TMA5 Removed	0.909856631	0.53761544	0.806060606	0.611270264

Table 40: Evaluation Results for MLP Experiments with Course A Data and hidden layer size = 250

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.922708653	0.653297258	0.815183983	0.699013881
Student Information Only	0.853507	0	0	0
Grades Only	0.917869944	0.591782107	0.847420635	0.672747442
Gender Removed	0.917895545	0.606706349	0.823888889	0.674435564
Age Removed	0.92109575	0.653297258	0.801111111	0.69309813
Region Removed	0.917869944	0.63281746	0.796944444	0.68675053
Age, Gender and Region Removed	0.921121352	0.601782107	0.858611111	0.676018118
TMA2 – TMA5 Removed	0.864797747	0.158636364	0.573333333	0.236295371
TMA3 - TMA5 Removed	0.885714286	0.378564214	0.785833333	0.469049736
TMA4 - TMA5 Removed	0.897004608	0.423185426	0.841666667	0.523045778
TMA5 Removed	0.908294931	0.506453824	0.81	0.593623277

Table 41: Evaluation Results for MLP Experiments with Course A Data and hidden layer size = 500

As with the Course A data set, changing the hidden layer size had minimal effect on the performance scores in any of the metrics when carrying out experiments using the Course B data. The effects of features are similar with good Recall scores for the most part, even when using only the earliest available grades, while Precision was heavily reliant on important features. The training times for this data set using Multi-Layer Perceptron were significantly higher than any other data set with any other model with the longest times being nearly a minute.

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.872273425	0.7101028	0.850798872	0.769798692
Student Information Only	0.667254408	0.168236055	0.563996784	0.25460796
Grades Only	0.881972582	0.709546293	0.873410599	0.781426446
Gender Removed	0.878198273	0.704786218	0.863723335	0.773855245
Age Removed	0.875509376	0.70998991	0.854509296	0.772124576
Region Removed	0.872540069	0.727676865	0.827786332	0.771208216
Age, Gender and Region Removed	0.879003275	0.728861491	0.852071954	0.78191185
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.760710663	0.33695002	0.716415755	0.451575214
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.78631206	0.43227985	0.761130187	0.543548351
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.827268643	0.551833853	0.822107132	0.6528426
TMA5 – TMA6 and CMA5 Removed	0.85475321	0.621316744	0.858481571	0.71648194

Table 42: Evaluation Results for MLP Experiments with Course B Data and hidden layer size = 100

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.881432774	0.691426634	0.892553707	0.774956704
Student Information Only	0.66675063	0.191505312	0.551660154	0.28190268
Grades Only	0.87496522	0.708607326	0.854927548	0.771649148
Gender Removed	0.873618236	0.719914396	0.841156148	0.772003893
Age Removed	0.871999536	0.730160915	0.82439914	0.772680394
Region Removed	0.884398458	0.714037766	0.879484903	0.786145789
Age, Gender and Region Removed	0.880354607	0.719771812	0.862174635	0.781375953
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.75558792	0.354382048	0.696048913	0.462768548
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.791434803	0.440487294	0.772384143	0.553424362
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.826467988	0.562719909	0.816570908	0.654427858
TMA5 – TMA6 and CMA5 Removed	0.853677941	0.665690752	0.822826707	0.729831456

Table 43: Evaluation Results for MLP Experiments with Course B Data and hidden layer size = 250

Data	Accuracy	Precision	Recall	F1-Score
All Features	0.876043388	0.746851506	0.831099328	0.78246152
Student Information Only	0.662468514	0.185031219	0.544154079	0.270015698
Grades Only	0.869305568	0.742873123	0.828317389	0.777306791
Gender Removed	0.883857925	0.712434828	0.876659246	0.784741027
Age Removed	0.867153581	0.766441369	0.791765558	0.773244536
Region Removed	0.879278613	0.703865118	0.871548741	0.775950034
Age, Gender and Region Removed	0.88008579	0.701039108	0.877827522	0.776395251
TMA2 – TMA6 and CMA2 - CMA5 Removed	0.760709214	0.36827274	0.688463675	0.476146105
TMA3 – TMA6 and CMA3 - CMA5 Removed	0.783346376	0.445704944	0.739244542	0.546171557
TMA4 – TMA6 and CMA4 - CMA5 Removed	0.824312379	0.538014675	0.830323296	0.641442208
TMA5 – TMA6 and CMA5 Removed	0.848287105	0.633902681	0.837528127	0.713107163

Table 44: Evaluation Results for MLP Experiments with Course B Data and hidden layer size = 500

The tuning parameter setting used in this experiment resulted in no real change in the performance with either data set. The only significant performance differences were seen when features were

removed. This is the slowest algorithm so far with good, but not outstanding, results in most configurations.

4.2. Discussion

4.2.1 Feature Selection

Having tested how different combinations affect machine learning model performance across all of the selected algorithms with a variety of parameter settings.

The removal of features which were found to have lower importance or were not correlated with the final outcome was investigated first. Removing individual features resulted in almost no change or varying changes in almost all situations. Gender, being poorly correlated and unimportant, was expected to have a greater effect when removed however, the change was never significantly more than for the removal of any other student information feature. Removing all student information features mostly resulted in a small negative performance hit. From these results it would seem that the removal of the least important features would only be beneficial or worthwhile if it was required in order to improve the time taken to train or if more relevant features are available as these would further diminish the importance of those already of low importance.

The use of student personal information as the only features cannot reliably produce accurate predictions with any of the algorithms and parameter configurations used in this study. Some form of relevant grade information is essential in order to develop a usable model.

In situations where not much grade information is available there was a clear and expected drop in performance with all algorithms and configurations. It is clear that the further into a class or degree a student is, the more relevant information will be available and the more accurately the prediction of their final outcome is likely to be. In the experiments carried out, this improvement in classification performance seems to saturate about halfway through single semester class, with a full undergraduate degree worth of classes, and using only the final grades for those classes, it is likely that the final degree outcome could be predicted with good accuracy within the first year due to significantly more relevant features being available. With a single year postgraduate degree it is likely, given these results, that after a single semester, the final degree outcome could be accurately predicted. This would be improved by using assessment performance throughout the first semester as predictors.

As should be expected, as more features with high importance or relevance are added the increase in performance for each additional feature decreases. It may be important to note the point at which very little improvement is seen and begin to determine which of the earlier grades being used are losing importance as these are likely to lose some relevance in longer degree programs though this would require further investigation out with the scope of this study.

4.2.2 Machine Learning Models

The Gaussian Naïve Bayes algorithm is clearly very heavily dependent on the data set available. While the resilience of the Recall score when using small numbers of features is advantageous, larger data sets are required to improve performance to a level which would be viable. This would likely be possible for popular, well established undergraduate courses but performance would be too poor to be workable in smaller courses such as post graduate masters programs. In the event that all of the courses for which predictions are being made have large training data sets available then Gaussian Naïve Bayes would be an ideal algorithm due to the simplicity and lack of any need for tuning.

K-Nearest Neighbours performed very well across all evaluation metrics with both data sets in most situations. Alongside this, it has very easy to tune parameters and has very short training times. The tuned algorithm is relatively resistant to changes in data set size and shape, able to handle large and small data sets as long as the features used are relevant. In the situation where the available data is changing or where more features are being added, e.g. as the year progresses and student performance is recalculated, the overall performance of K-Nearest Neighbours is likely to remain high. When correctly tuned K-Nearest Neighbours can perform well with even a small number of features making it ideal for early student performance prediction. The short training and testing times would also make it viable for further development of an application for predicting individual student performance. Of the five algorithms investigated in this study this was the easiest and most intuitive to set up and tune. It also provided some of the best results for prediction when considering Recall score as the most important metric though the Precision scores with smaller data sets and less features are very poor. K-Nearest Neighbours would work well for well established undergraduate and postgraduate degrees for which there is plenty of data available from previous cohorts for training.

Similar to K-Nearest Neighbours, Random Forest performed well with all evaluation metrics in both data sets and with most parameter configurations. Random Forest does mostly benefit from the removal of the least important features which is unsurprising as the Random Forest feature importance functionality was used to determine the importance of the features. With more grades added the Random Forest performance extremely well and it can be assumed that this would be the case when using grade data from a full degree program. Where Random Forest falls short is with smaller data sets combined with a small number of features, this results in performance dropping off relatively sharply. For this reason Random Forest would only really be suitable for full undergraduate degree prediction where there is likely to be more past data available and more grades throughout the year to use as predictors. Where K-Nearest Neighbours has the best Recall scores, Random Forest has the best Precision scores. This supports the suggestion that it would be best suited for undergraduate degrees as these tend to have more students which can increase the cost of misclassifying passing students as failing students.

When fully tuned to maximise performance Support Vector Machines seem to be excellent classifiers for the purposes of identifying struggling students. Unfortunately, it is getting to this point which is most difficult and unclear for SVM as a poorly tuned model can result in extreme overfitting as can be seen in the results using high Gamma, especially those on the Course B data set. The data set used can drastically change the values for optimal tuning which makes tuning more involved than for other algorithms. There are methods available with Scikit Learn for automatically tuning parameters which were not explored in this study however, they are known to be extremely computationally expensive and so would only really be suitable in a situation where the model is trained once before running a large number of predictions. For a university predicting student performance across a large number of degrees where each degree has its own large data set which would require identifying the best tuning parameter values or where those data sets frequently change this would likely be a poor solution. A more focussed study looking at the automatic tuning of the SVM parameters and optimisation of the data used may help to alleviate some of the issues found in this study.

The experiments carried out using Multi-Layer Perceptron did not yield particularly good results. The tuning parameter configuration gave no real variation in any of the performance metrics and the performance differences seen between experiments carried out on each data set were not significant. The main benefit seems to be that the Recall score remained low even with a small number of important features which mean that even early in a course or degree this classifier would perform well. Despite this, most of the other algorithms investigated in this study performed better than this

one and would be better suited for predicting student performance. The training time is also considerably higher for Multi-Layer Perceptron however, this is unlikely to be a deciding factor, especially in the situation where predictions are periodically carried out for an entire cohort. If an application was to be designed to make predictions on a student by student basis this would likely be a poor candidate.

5. Recommendations and Conclusions

5.1 Conclusion

The main objectives of this study were:

- To determine if machine learning models were viable as a means of identifying students at risk of failing.
- To determine how early in a course is it possible to accurately make these predictions.
- To compare a range of machine learning algorithms and identify those which may be suitable for this purpose.
- To determine how differences in data sets affect the performance of machine learning algorithms in an academic context.
- To determine whether features other than student grades are beneficial in making predictions.

Through a series of experiments on two data sets using multiple classification algorithms with a variety of parameter configurations it can be concluded that, with the correct algorithm and optimised tuning parameters, machine learning is an extremely useful tool in predicting student performance, capable of being adapted to any course as long as there is sufficient data available from previous cohorts. The two algorithms which stood out as viable candidates were Random Forest and K-Nearest Neighbours, both of which performed well, are simple to set up and tune and are very quick to train even when using large data sets.

The data available plays a significant role in the overall performance of any model. A larger pool of data will provide more training examples but will also allow for flexibility in culling outlying data in an effort to balance the examples from each of the classification outcomes. This will provide further improvements to the overall performance of the trained model. It has been shown that the features to be considered should be limited to earlier grades achieved when there are several available. In the event that not enough grades are available then limiting further features to those likely to affect a student's academic career is best. In this study it was possible in some cases to achieve a good performance with around half of the assessment marks attained in a single semester. In a full degree program significantly more relevant grade data will be available and it can be reasonably assumed that this will be sufficient to provide accurate predictions as early as a year into a four year undergraduate degree program or a semester into a single year postgraduate masters program.

The experiments carried out in an effort to determine what student information was relevant to prediction determined that, while the addition of certain information such as previous educational achievements or IMD band can sometimes have a positive effect on the overall performance of a model, any contribution is ultimately overshadowed by the far greater relevance of any previously achieved grades. Student personal information cannot stand as the only predictors, any predictions made without relevant grades will be extremely unreliable. As mentioned, in any full degree program there will be enough relevant grade information that using student personal information would be detrimental to the performance of any classifier.

5.2 Recommendations for Future Work

One of the greatest limitations faced during this study was the available data sets. The original plan for the study was to use a data set consisting of data taken from Strathclyde University students who had previously finished a full degree programme in an effort to predict degree outcomes. Unfortunately it was discovered at a very late stage that this data could not be made available and as such the Open University Learning Analytics data set, which consisted of student data from individual

anonymised courses, was selected for use instead due to its ready availability. For this reason, it would be extremely beneficial for a similar study to be carried out on a variety of full degree programs to provide more conclusive evidence for the effectiveness of machine learning in predicting student performance. By selecting different degree programs, types of degree programs and different sizes of those degree programs more information could be gleaned on which of the classifiers investigated is most effective in most situations.

Another limitation of this study is that the broad range of algorithms investigated made it difficult to explore the full range of tuning parameters available to the algorithms found to be most effective, in this case Random Forest and K-Nearest Neighbours. With the use of data sets for full degree programmes it would also be beneficial to investigate the effectiveness of tuning this smaller selection of algorithms. With the smaller selection of algorithms, it would be possible to comprehensively investigate how best to tune with different data sets and provide further evidence as to which algorithm is most fit for this purpose.

References

Adamson, J., & Clifford, H. (2002). An Appraisal of A-Level and University Examination Results for Engineering Undergraduates . *International Journal of Mechanical Engineering Education, 30*(3), 265-279.

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & Costa, L. d. (2014). A Systematic Comparison of Supervised Classifiers. *PLoS ONE, 9*(4).

Anderson, T., & Anderson, R. (2017). Applications of Machine Learning to Student Grade Prediction in Quantitative Business Courses. *Global Journal of Business Pedagogy, 1*(3), 13+.

Birch, D. M., & Rienties, B. (2014). Effectiveness of UK and international A-level assessment in predicting performance in engineering. *Innovations in Education and Teaching International, 51*(6), 642-652.

Chee, K. H., Pino, N. W., & Smith, W. L. (2005). Gender differences in the academic ethic and academic achievement. *College Student Journal, 39*(3), 604-618.

Cunningham, P., & Delany, S. J. (2007). k-Nearest Neighbour Classifiers. *Technical Report UCD-CSI-2007-4*. Dublin: Artificial Intelligence Group.

Day, I. N., Blankstein, F. M., Westenberg, M., & Admiraal, W. (2018). A Review of the Characteristics of Intermediate Assessment and their Relationship with Student Grades. *Asessment & Evaluation in Higher Education, 43*(6), 908-929.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research 3*, 1289-1305.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 2627-2636.

Hall, M. A., & Smith, L. A. (1999). Feature Selection for Machine Learning: Comparing aCorrelation-based Filter Approach to the Wrapper. *Proceedings of the Twelfth International Florida Artificial Intelligence Research Symposium Conference*, (pp. 235-239). Orlando, Florida.

Jacob, S. G., Sridhar, S., & Murugavel, N. (2017). A Comparative Study on the Performance of Classifiers in Predicting Frequency of Drug Intake: A case study with Ketamine, Heroin, Crack and Meth. *International Journal of Engineering Technology Science and Research, 4*(9).

Khalid, S., Khalil, T., & Nasreen, S. (2014). A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. *2014 Science and Information Conference*, (pp. 372-378). London, UK.

Kira, K., & Rendell, L. A. (1992). Feature Selection Mehods and Algorithms. *AAAI'92 Proceedings of the Tenth National Conference on Artificial Intelligence*, (pp. 129-134). San Jose, California.

Lemus-Zúñiga, L. G., Montañana, J. M., Buendia-Garcia, F., Poza-Luján, J. L., Posadas-Yagüe, J. L., & Benlloch-Dualde, J. V. (2015). Computer-assisted method based on continuous feedback to improve the academic achievements of first-year students on computer engineering. *Computer Applications in Engineering Education, 23*(4), 610-620.

Liu, C., Kubler, S., & Yu, N. (2014). Feature Selection for Highly Skewed Sentiment Analysis Tasks. *Proceedings of the Second Workshop on Natural Language Processing for Social Media*, (pp. 2-11). Dublin, Ireland.

Maclin, R., & Opitz, D. (1997). An Emprical Evaluation of Bagging and Boosting. *Proceedings of the Fouteenth National Conference*, (pp. 546-551). Providence, Rhode Island.

Shaw, S., & Bailey, C. (2011). An American university case study approach to predictivevalidity: Exploring the issues. *Research Matters*(12), 18-26.

Tan, A. C., & Gilbert, D. (2003). An empirical comparison of supervised machine learning techniques in bioinformatics. *APBC '03 Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics , 19*, 219-222.

Teodorovic, J. (2012). Student background factors influencing student achievement in Serbia. *Educational Studies*, 89-110.

Tieben, N., & Wolbers, M. (May 2010). Success and failure in secondary education: socio-economic background effects onsecondary school outcome in the Netherlands, 1927–1998. *British Journal of Sociology of Education, 31*(3), 277-290.

Zhang, N., & Henderson, C. N. (2015). Can formative quizzes predict or improve summative exam performance? *The Journal of Chiropractic Education, 29*(1), 16-21.