Honours Research

Automatic Labelling of Student Results as PCD, RET, MBR and MRNM

Jettiniel Chepiri
734438

Supervised by Ritesh Ajoodha and Benjamin Rosman

November 13, 2017

Abstract

The ability to predict a student's performance could be useful in a great number of different ways. This presents an approach to create an auto-tagging system which examines student grades and assigns an outcome-code. The system will attempt to learn the delicate boundaries between these class labels that depict academic performance.Our auto-tagging system can provide a meaning evaluation of student performance based only on the results specified.

Keywords: predicting student performance, statistical prediction, supervised learning

Declaration

I, Jettiniel Chepiri, hereby declare the contents of this Research proposal to be my own work. This Research is submitted for the degree of Bachelor of Science with Honours at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.

Acknowledgements

I would like to thank my supervisors, Ritesh Ajoodha and Benjamin Rosman, for the support and the motivation he gave me during my research. I would like also to thank my Friends who are always there for me.

Contents

List of Figures

Chapter 1

Introduction

For higher education institutions whose goal is to contribute to the improvement of quality of higher education, the success of creation of human capital is the subject of a continuous analysis. Therefore, the prediction of student's success is crucial for higher education institutions, because the quality of teaching process is the ability to meet students' needs [Al-Radaideh *et al.* 2006]. The sole purpose of this research is to develop a system that can classify student results into one of the following possible out-come codes include PCD, where the student can proceed to the next year of study; RET, where the student needs to repeat some subjects in order to continue to the next year; MBR, which indicates that the student will be excluded for poor performance but does not need to appeal to be allowed re-admission and MRNM, which indicates that the student has been excluded and will have to appeal for re-admission [Minaei-Bidgoli *et al.* 2003]. By converting data into knowledge, the gratification of all participants is attained: students, professors, administration, supporting administration, and social community [Kotsiantis 2012].

Some similar systems have been already implemented in other countries to facilitate the classification of students according to their performance [Kotsiantis 2012]. The system is really crucial in the society as it helps students to save money, if a students is classified as MBR she can change the courses and try to do other courses she might be good at. Many students get enrolled in universities without proper understanding of the courses they are doing hence they will experience difficulties in understanding the concept of their subjects, this system will help to identify these kind of students and they will be given proper assistance [Al-Radaideh *et al.* 2006].

1.1 Problem statement and Motivation

Most universities have young lectures who does not have enough experience to understand properly how to use previous students performance to make decisions, they can use this system to to make predictions about students performance. Younger academics don't usually have knowledge of previous students results classification.

The target of the system is mostly for education institution like universities and colleges. For instance, in the University of the Witwatersrand most of the students fail first year and repeat several times before they graduate. With high rate of drop out in south African in higher education sector, this system is a solution to a national crisis and Implementing it our education institutions can save resources. This will be the first expect system that is able to predict whether a student is likely to success or fail in a certain course given students academic history at Wits university [Kotsiantis 2012].

1.2 Hypothesis

Our research will make use of machine learning algorithms to to classify our data. From a general point of view, the problem can be reduced to this question: Given student previous marks can we predict the out-come code of a student? The Wits University has designed a set of outcome codes which is intended to classify a students academic report into one of four possible outcome codes, possible outcome codes include PCD where the student can proceed to the next year of study; RET, where the student needs to repeat some subjects in order to continue to the next year MBR, which indicates that the student will be excluded for poor performance but does not need to appeal to be allowed re-admission and MRNM, which indicates that the student has been excluded and will have to appeal for re-admission. The aim of our research is to train a machine learning classifier to make a difference between a students in different outcome codes and return the result [Minaei-Bidgoli *et al.* 2003].

1.3 Contribution

This research will investigate to make a system that will auto-tagging system which examines student grades and assigns an outcome-code. Your system will attempt to learn the delicate boundaries between these class labels (PCD, RET, MBR, and MRNM) by accessing training data assembled, over many years, by the teaching and learning committee. Such committees can use the system to assist them with making decisions on how to classify student grades.Recent technologies and previous data can be gathered together to make decision easier [Kotsiantis 2012]. Current thinking is well structured and founded, my research doesn't discount the previous theories but rather build upon them. This paper bring together into one solid ideology that has developed over years and my purpose is bridge the gap buy adding new thing on the research that already been cover most the staff. This research focus on a Witwatersrand university and l will use Wits data to build my model. My area of focus will be to cover the areas not covered by other publisher. This is the first time research of this nature to be done at Wits [Minaei-Bidgoli *et al.* 2003].

The remainder of the proposal is structured as follows, chapter 2 presents the work done by other researchers with regard to the students classification. It also deals with the background of the project and also compare many literature that focus on predicting students performance. Chapter 3 describes the two main methods we are going to use in our classification. The research hypothesis is stated in Section 3.2 and the methodology that is used has been presented in Section 3.3. Chapter 4 give the Results for the research and analysis as well. Chapter 5 provides a summary of the entire research paper.

Chapter 2

Background and Related Work

2.1 Introduction

The system is easy to read and understood. This system can give professor interesting information about student and provides guidance to lectures to choose a suitable track, by analyzing experiences of students with similar academic achievements. Without investing much in understanding the academic trends is the reason why universities have high drop out rate [Yadav and Pal 2012]. In this research we will attempt to create an auto-tagging system which examines student grades and assigns an outcome-code.The system will attempt to learn the delicate boundaries between these class labels (PCD, RET, MBR, and MRNM) by accessing training data assembled over many years. In this chapter, we review literature work that has been done in the field of predicting the students by different scholars before. The following work covers a broad range of variables that affect the performance of a student.

2.2 Methods in different Literature

Many theories has been proposed to explain how we predict students performance.There are many different classifiers in the literature and one cannot choose the best, because they differ mutually in many aspects such as: learning rate, amount of data for training, classification speed, robustness, etc. Although the literature covers a wide variety of such theories depending on different scholars, the algorithm below are the most used ones.

- Support vector machine.
- Mining student data using decision trees
- Bayesian network algorithms, Naive Bayes
- Random forest
- Linear regression Model
- Multi-layer Perceptions

Although the literature present these themes in a variety of context, this research is going use Support vector machine, Randomforest, Classification via Regression, Naive Bayes and Multilayer Perception.

As stated in the introduction, the input data used in the research is from Wits university. Wits a good sample it can resemble many university in South Africa and in other countries as well . In our case, many universities, colleges and high school they can use the same model when we change our training data and features. Below we are giving an outline of what other scholars focused on.

2.3 Predicting the outcome code of students

2.3.1 Using Classification to predict student performance

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. The scope of this survey is restricted to comparing some popular non-parametric pattern classifiers and a single parametric pattern classifier according to the error estimate. By reading though many different literature there are six different classifies l have come across with above classifier in addition to ANN neighbor KNN nearest neighbor [Yadav and Pal 2012].

2.3.1.1 Combining classifiers

Many publishers use a single classifier to classify their data, this has a lot of limitation to level of accuracy hence it affect the decision making for example used decision tree as their only classifier wish minimized the accuracy of their results. In such cases it would be better to pool the results of different classifiers to achieve the optimal accuracy. Every classifier operates well on different aspects of the training or test feature vector[Bunkar *et al.* 2012]. As a result, assuming appropriate conditions, combining multiple classifiers may improve classification performance when compared with any single classifier.By combining multiple classifiers we hope to improve classifier performance. After some prepossessing operations were made on the dataset, the error rate of each classifier is reported. finally, to improve performance, a combination of classifiers is presented.There are different ways one can think of combining classifiers.

2.3.1.2 Ways combining classifiers

- The simplest way is to find the overall error rate of the classifiers and choose the one which has the least error rate on the given dataset. In general, it has a better performance than individual classifiers.
- The second method, which is called online CMC, uses all the classifiers followed by a vote. The class getting the maximum votes from the individual classifiers will be assigned to the test sample. A combination of multiple classifiers leads to a significant accuracy improvement in all cases, many publisher of journals overlooked the importance of using multi-classifiers [Bunkar *et al.* 2012].

2.3.1.3 Difference of Methodologies

Many of the references used different source of data, some they analyzed online students based on the web based system. Some publisher did not take into account many factors that can affecting student

performance and this affect conclusion of the paper [Minaei-Bidgoli *et al.* 2003]. Different publisher used different ways to estimate parameter and they use assumption that may lead to different outcome.

2.4 Contribution

This is the first system to access students and provide prediction about students performance. This research bring together into one solid ideology that has developed over years and my purpose is bridge the gap buy adding new thing on the research that already been cover most the staff. My area of focus will be to cover the areas not covered by other publisher. This is the first time research of this nature to be done at Wits.

2.5 Conclusion

Current thinking is well structured and founded, my research doesn't discount the previous theories but rather build upon them. This paper bring together into one solid ideology that has developed over year and my purpose is bridge the gap buy adding new thing on the research that already been cover most the staff. However predicting students performance dependence on a particular school and in my research l will focus on a Witwatersrand university.

Chapter 3

Research Methodology

3.1 Introduction

Classification of students and predicting their outcome code as stated previous chapter is very challenging and machine learning techniques have been proven to work better than other techniques [Aksenova *et al.* 2006].This chapter will describe the hypothesis and the research question about the current problem. The hypothesis and the research question are defined and presented in Section 3.2. The Section 3.3 will focus on the methodology used for how the data will be obtained and processed to test the hypothesis and thereafter answer our research question [Al-Radaideh *et al.* 2006]. That section is broken into some sub-parts to accept our hypothesis. In subsection 3.3.1, we briefly present how the data was collected [Bunkar *et al.* 2012]. The subsection describes how the data was prepared for the data labeling in subsection. This section describes how the data was labeled in order to create categories of the students marks depending on their courses that will use in the training and testing subsection.

3.2 Research Hypothesis

Nowadays, at Witwatersrand university students outcome codes is becoming a big issue as stated in Chapter 1.Failure to perform this role skilfully will result in a loss of potential talent that the university could have later benefited from [Kotsiantis 2012].Our research will focus on a very specific angle on an approach to a solution to the problem of finding auto-tagging system which examines student grades and assigns an outcome-code: **how can some automated techniques like machine learning bring a solution to the problem of classifying students whether they will pass on not?** Based on that question we can give the following hypothesis: **It can be predicted that students marks can be classified at possible outcome codes include PCD where the student can proceed to the next year of study; RET, where the student needs to repeat some subjects in order to continue to the next year MBR, which indicates that the student will be excluded for poor performance but does not need to appeal to be allowed re-admission and MRNM, which indicates that the student has been excluded and will have to appeal for re-admission.**

3.3 Methodology

3.3.1 Data Collection

The data was collected from the University of the Witwatersrand, academic information system unit(AISU). The data was a set of marks which show students performance thought the period of study at the University of the Witwatersrand. The selected data has been chosen because it corresponds to a good sample that is identical to other good university in the world . Before performing any process on the data, the settings of the course of which the student should be taken into consideration so that we do not use poor type of data to perform our classification.

3.3.2 Data Preparation

For this step, the collected data were prepared in the table format that is suitable for used in data mining system. The data was cleansed by removing the various inconsistent values using the same standard value using the majority data approach. Since the collected data may have some irreverent attributes that may degrade the performance model , a feature selection approach was done. For this purpose the WEKA toolkit was used to eliminate by feature selection approach [Al-Radaideh *et al.* 2006].

3.3.2.1 Datasets

The data being labeled, one dataset was created for each of the training and the testing phase respectively. Those datasets was generated from the output of the data labeling section. Four main categories was generated from the data labeling:

- **PCD,** which will contain categories of students can proceed to the next year of study
- **RET,** which will contain a categories of students needs to repeat some subjects in order to continue to the next year
- **MBR,** which will contain a categories of students will be excluded for poor performance,but does not need to appeal to be allowed re-admission
- **MRNM ,**which will contain a categories of students which been excluded and will have to appeal for re-admission.

3.3.3 Out-tagging Outcome code

Lecture will use the system by simply following easy procedures, after marking the students scripts lecture will enter students marks into the system, the system will automatically input that data into our algorithm an output code will appear on the students self service account indicating one of the classes above.

3.3.3.1 Server Side

The server will alliteratively check for update on new data so that the parameter of our algorithms will change with time. That file will be stored in the server as a dataset. Every time the system received a new data set the parameter can change in-order to improve the prediction in the future.

3.4 Types of methods

a) Support Vector Machines

b) Naive bayes

c) RandomForest

d) Classification Via Regression

e) MultilayerPerceptron

3.4.1 Support vector Machine

Support Vector Machine is a classification technique that is listed under supervised learning models in Machine Learning. It involves finding the hyper-plane that best separates two classes of points with the maximum margin. Essentially, it is a constrained optimization problem where the margin is maximized subject to the constraint that it perfectly classifies the data. The data points that kind of "support" this hyper-plane on either sides are called the "support vectors" [Romero and Ventura 2007]. A problem involving multiple classes can be broken down into multiple one-versus-one or one-versus-rest binary classification problems.

Given training vectors $x_i \in \mathbb{R}^p$, i=1,..., n, in two classes, and a vector $y \in \{1, -1\}^n$, SVC solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$
$$\zeta_i \geq 0, i = 1, ..., n$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$
$$\text{subject to } y^T \alpha = 0$$
$$0 \leq \alpha_i \leq C, i = 1, ..., n$$

where e is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The decision function is:

$$\text{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho)$$

Figure 3.1: Support vector machine equations

How do we choose support vectors ? It is actually a by-product of finding weights (the alphas). So in training phase, we find weight for each training point, and those points whose weight becomes zero are not support vectors i.e. their importance is zero during test time, and rest are support vectors [Romero and Ventura 2007].

So then how to learn alphas ? It is at this point, where all the math comes. Intuitively, we try to find that separating hyperplane, from which distance of closest training points is maximum (also known as max-margin classifier), and those closest training points then become support vectors [Aksenova *et al.* 2006]. During the math, while optimizing the function, we get the weights of the support vectors.Now once we have selected support vectors, we assign a weight to each support vector, which basically tells how much importance we want to give to that support vector while making our decision.

3.4.1.1 Advantages of support vector machines

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

3.4.1.2 Disadvantages of support vector machines

- If the number of features is much greater than the number of samples, the method is likely to give poor performances.
- Support vector machine do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

Support vector machines focus only on the points that are the most difficult to tell apart, whereas other classifiers pay attention to all of the points. Unlike other classifiers, the support vector machine is explicitly told to find the best separating line. How? The support vector machine searches for the closest points (Figure 3.1), which it calls the ”support vectors” (the name ”support vector machine” is due to the fact that points are like vectors and that the best line ”depends on” or is ”supported by” the closest points) [Aksenova *et al.* 2006].

3.4.2 Naive Bayes classifier

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors.In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [Romero and Ventura 2007]. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as Naive[Aksenova *et al.* 2006].Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to.outperform even highly sophisticated classification methods.Bayes theorem provides a way of calculating posterior probability P(c—x) from P(c), P(x) and P(x—c). Look at the equation below:

- P(c—x) is the posterior probability of class (c,target) given predictor (x,attributes).
- is the prior probability of class.

- P(x—c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Figure 3.2: Bayes equation

3.4.2.1 Advantages of Naive Bayes

- It is easy and fast to predict class of test data set. It also perform well in multi-class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variables. For numerical variable, normal distribution is assumed.

3.4.2.2 Disadvantages of Naive Bayes

- On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict proba are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

- If categorical variable has a category in the test set, which was not observed in training data set, then model will assign a zero probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique [Bunkar *et al.* 2012].

3.4.2.3 Ways improve the power of Naive Bayes Model

- If continuous features do not have normal distribution, we should use transformation or different methods to convert it in normal distribution.
- If test data set has zero frequency issue, apply smoothing techniques Laplace Correction to predict the class of test data set.
- Remove correlated features, as the highly correlated features are voted twice in the model and it can lead to over inflating importance.

3.4.3 RandomForest

Random forest algorithm is a supervised classification algorithm. It creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest [Romero and Ventura 2007].

3.4.3.1 Advantages of Naive Bayes

- Random forest classifier will handle the missing values
- random forest classifier wont overfit the model, In case we have more trees.
- It can use for both classification and the regression task.

3.4.4 MultilayerPerceptron

Multilayer perceptron classifier (MP) is a classifier based on the feedforward artificial neural network. MP consists of multiple layers of nodes. Each layer is fully connected to the next layer in the network. Nodes in the input layer represent the input data [Aksenova *et al.* 2006]. All other nodes map inputs to outputs by a linear combination of the inputs with the nodes weights and bias and then apply the activation function. Figure 3.3 show the layers.

Hidden layers are in between input and output layers. Typically, the number of hidden layers range from one to many. It is the central computation layer that has the functions that map the input to the output of a node. The output layer is the final layer of a neural network that returns the result back to the user environment [Bunkar *et al.* 2012].

3.4.5 ClassificationViaRegression

Class for doing classification using regression methods. Class is binarized and one regression model is built for each class value. The problem of multiclass classification is considered and resolved through

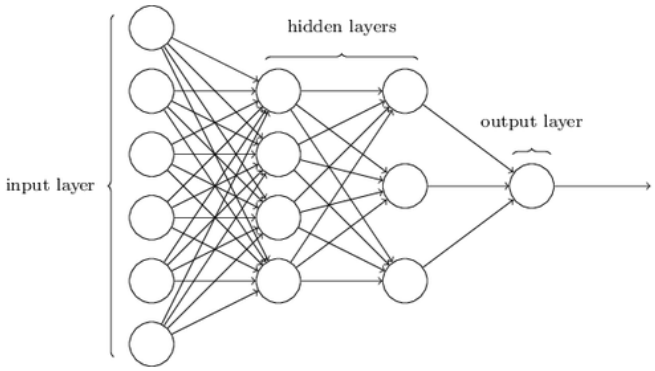

Figure 3.3: Multilayer Perceptron Architecture

the mul-tiresponse linear regression approach. Scores are used to encode the class labels into multivariate responses. The regression of scores on input attributes is used to extract a low-dimensional linear discriminant subspace [Aksenova *et al.* 2006]. The classification training and prediction are carried out in this low-dimensional subspace. A test point is classified to the nearest class centroid of fitted values in the measure of Mahalanobis distance. The multi-response linear regression can extend to a nonlinear one by the kernel trick. The regression approach provides a simple alternative for multiclass support vector classification.

3.5 Conclusion

This chapter discussed the methodology used in our research to support our stated hypothesis and brought a potential answer to our research question. we used five algorithm to perform our experiment and each algorithm is unique The next chapter will discuss and analyses the results we got from this research.

Chapter 4

Results

4.1 Introduction

The study main objective is to find out if it is possible to predict the class variable using the explanatory variables which are retained in the model. Several different algorithms are applied for building the classification model, each of them using different classification techniques. The WEKA Explorer application is used at this stage. Each classifier is applied for two testing options cross validation (using 10 folds and applying the algorithm 10 times each time 9 of the folds are used for training and 1 fold is used for testing) and percentage split (2/3 of the dataset used for training and 1/3 for testing). Below is the phases of the research.

Phase 1: Data Collection
Phase 2: Data Preparation
Phase 3: Training
Phase 4: Testing

4.2 Research's Phases

4.2.1 Phase 1: Data Collection

The data was collected from the Academic information and systems Unit Department of Wits university . I used first years marks for students under the faculty of science.

4.2.2 Phase 2: Data Preparation

In this phase, the data was cleaned and analyzed to produce the required datasets and features that l used for prediction. This phase was the longest and the most important part of the research. l deleted unwanted features and incomplete data. The data was also converted into afrr format that which is best to use in Weka. This phase was also the most sensitive because if the data is not well prepared, wrong results will be produced.Students marks dataset was work intensive because each column of the data must be labeled properly for the classifier to perform well otherwise the data will be mixed up.

4.2.3 Phase 3: Training

At this stage, a classifier as a function that is trained to arrange to some inputs into categories. In our case, that classifiers was able to classify a student in one the categories PCD, RET, MBR or MRNM depending on the current situation of the student performance. The training phase is the second longest phase in the process. All the parameters of the classifier was well defined to optimize the performance of the classifier.

4.2.4 Phase 4: Testing

The testing phase was very important to check whether our system works properly based on the requirements. A third of our data was used for testing.

4.3 Results and discussion

Unlike previous attempts related to grade prediction that focus on a single algorithm and do not perform any form of feature selection, the overall goal of this paper is to find the best combination of learning algorithms and selected features in order to achieve more accurate prediction in datasets with an balanced class distribution and a small number of instances. In this research we focus on Naive Bayes(NB), Support Vector Machine(SOM), ClassificationViaRegression(CVR), RandomForest(RF) and Multilayer Perceptron(MP). Using these techniques many kinds of knowledge can be discovered and used for the benefit of students .

4.3.1 Naives bayes

Below is the summary of results we get from Naive Bayes. Relatively low accuracy, 74percent of the of the students are correctly classified. Kappa is a chance-corrected measure of agreement between the classifications and the true classes. Kappa is greater than zero which means the model is doing better than chance.

Below is the confusion matrix for Naive Bayes, 41 students are correctly classified as MBR, 98 as MRNB, 228 as PCD, 61 AS RET. The rest are wrongly classified. The model can be improved by increasing our dataset.

```
a   b   c   d   e   f   g   <-- classified as
0   0   1   0   0   0   0 |   a = MBP
0  41   9   1   9   0   0 |   b = MBR
0  19  98   9   4   0   0 |   c = MRNM
0   3   6 228  37   0   0 |   d = PCD
0  28   5  13  61   0   0 |   e = RET
0   0   1   0   0   0   0 |   f = XXXX
0   0   0   0   0   0   0 |   g = ****
```

Figure 4.1: Confusion Matrix of Naive Bayes

4.3.2 Support Vector Machine

Below is the summary of results we get from Support Vector machine .Fair accuracy, 85percent of the of the students are correctly classified. Kappa is greater than zero which the means model is doing better than chance.

Below is the confusion matrix for Support Vector Machine, 38 students are correctly classified as MBR, 111 as MRNB, 263 as PCD, 80 AS RET. The rest are wrongly classified.

```
a   b   c   d   e   f   g   <-- classified as
0   0   1   0   0   0   0 |   a = MBP
0  38  12   3   7   0   0 |   b = MBR
0  10 111   7   2   0   0 |   c = MRNM
0   1   1 263   9   0   0 |   d = PCD
0   7   5  15  80   0   0 |   e = RET
0   0   1   0   0   0   0 |   f = XXXX
0   0   0   0   0   0   0 |   g = ****
```

Figure 4.2: Confusion Matrix/SVM

4.3.3 Random Forest

Below is the summary of results we get from Random forest classifier. Relatively high accuracy, 89.4percent of the of the students are correctly classified. Kappa is greater than zero which the means model is doing better than chance.

Below is the confusion matrix for Random Forest, 48 students are correctly classified as MBR, 113 as MRNB, 270 as PCD, 81 AS RET. The rest are wrongly classified..

```
a   b   c   d   e   f   g   <-- classified as
0   0   0   1   0   0   0 |   a = MBP
0  48   5   2   5   0   0 |   b = MBR
0   6 113   7   4   0   0 |   c = MRNM
0   0   1 270   3   0   0 |   d = PCD
0   3   5  18  81   0   0 |   e = RET
0   0   1   0   0   0   0 |   f = XXXX
0   0   0   0   0   0   0 |   g = ****
```

Figure 4.3: Confusion Matrix of Randomforest

4.3.4 MultilayerPerceptron

Below is the summary of results we get from Multilayer Perceptron classifier. Fair accuracy , 84.8percent of the of the students are correctly classified, performed better than Naive Bayes. Kappa is greater than zero means which the means model is doing better than chance.

Figure4.8 is the confusion matrix for MP, 40 students are correctly classified as MBR, 109 as MRNB, 253 as PCD, 84 AS RET. The rest are wrongly classified. The model can be improved by increasing the data set.

```
a   b   c   d   e   f   g   <-- classified as
0   1   0   0   0   0   0 |   a = MBP
0  40  11   2   7   0   0 |   b = MBR
0  12 109   6   3   0   0 |   c = MRNM
0   1   1 253  19   0   0 |   d = PCD
0   5   4  14  84   0   0 |   e = RET
0   0   1   0   0   0   0 |   f = XXXX
0   0   0   0   0   0   0 |   g = ****
```

Figure 4.4: Confusion Matrix of MultilayerPerception

4.3.5 ClassificationViaRegression

Below is the summary of results we get from ClassificationViaRegression. The accuracy is better than Nave Bayes ,MR and SMO , 87.4percent of the of the students are correctly classified. Kappa is greater than zero which the means model is doing better than chance.

Below is the confusion matrix for CVR, 47 students are correctly classified as MBR, 111 as MRNB, 262 as PCD, 81 AS RET. The rest are wrongly classified.

```
a   b   c   d   e   f   g   <-- classified as
0   0   0   1   0   0   0 |   a = MBP
0  47   6   1   6   0   0 |   b = MBR
0   5 111   9   5   0   0 |   c = MRNM
0   0   2 262  10   0   0 |   d = PCD
0   6   4  16  81   0   0 |   e = RET
0   0   1   0   0   0   0 |   f = XXXX
0   0   0   0   0   0   0 |   g = ****
```

Figure 4.5: Confusion Matrix of CVR

4.4 Comparison and analysis

Comparing results above we can see that some algorithms performs better, towards this goal, we compared accuracy of five different ML algorithms applied to students previous marks. Maximum accuracy was achieved at 89.35pct for Random Forest, followed by 87.43pct for Classificationbyregression, 85pct for Support Vector Machine, 84.81pct for a MultilayerPerception and 74.69pct for Naive Bayes. As we can see from the graph below Random forest is the highest meaning is the best classifiers than the other algorithms.

Figure4:12 is the diagram shows the relative absolute error for all algorithms, Naives bayes is our test base. Support vector machine is significantly better than Naive Bayes, the rest are significantly poor compared to Baive Bayes.

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4 -D 1 -R 2 -S 0.05 -result-matri
Analysing:  Relative_absolute_error
Datasets:   1
Resultsets: 5
Confidence: 0.05 (two tailed)
Sorted by:  F_measure
Date:       11/12/17 10:46 AM


Dataset                   (1) bayes.Na | (2) functi (3) trees (4) funct (5) meta.
----------------------------------------------------------------------------------
cs-I-training             (10)   39.14 |   106.66 v   30.12 *   24.64 *   33.18 *
----------------------------------------------------------------------------------
                               (v/ /*) |    (1/0/0)   (0/0/1)   (0/0/1)   (0/0/1)


Key:
(1) bayes.NaiveBayes
(2) functions.SMO
(3) trees.RandomTree
(4) functions.MultilayerPerceptron
(5) meta.ClassificationViaRegression
```

Figure 4.6: Relative absolute errors

ROC is one of the most important values output by Weka, it gives you an idea of how the classifiers are performing in general. Since ROC is greater than o.5 of all the algorithm we can say all our algorithms perform well on average, with Random Forest being the best compared to others. Trying to take a deeper look at the obtained results, it is presented in our weka results detail that f-measure for all class values in all our datasets, although the Random forest is the only the learning algorithms with predominately best results.

4.5 Conclusion

Comparing the results above, it is evident that Random Forest is the best Classifier. An interesting finding from this results show that most our models accurately predict the students final performance, given our students datasets especially if we do good feature selection.The overall prediction accuracy in our analysis varies from 75pct to 89pct. Results are more than promising and we are able to do future implementation of a student performance prediction tool for many universities.

Chapter 5

Conclusion

From this research paper we learn the important of predicting students marks using supervised machine learning methods. We created a system that we can use to predict students out-come codes, it will help lecture to mark decision whether the student should proceed or not. The data was collected from Wits university and it was grouped into training and testing data.Comparing we got it is evident that Random Forest is the best Classifier. An interesting finding from this results show that our models accurately predict the students final performance, given our students datasets especially if we do good feature selection.The overall prediction accuracy in our analysis varies from 75pct to 89pct.

The Wits University has designed a set of outcome codes which is intended to classify a students academic report into one of four possible outcome codes Possible outcome codes include PCD where the student can proceed to the next year of study; RET, where the student needs to repeat some subjects in order to continue to the next year MBR, which indicates that the student will be excluded for poor performance but does not need to appeal to be allowed re-admission and MRNM, which indicates that the student has been excluded and will have to appeal for re-admission. The user of the system will simply input student marks and the out-come code will be out-tagged.It can be concluded that this methodology can be used to help students and lectures to improve students performance, reduce failing ratio by taking appropriate steps at right time to improve the quality of learning.

Current thinking is well structured and founded, my research doesn't discount the previous theories but rather build upon them. This paper bring together into one solid ideology that has developed over year and my purpose is bridge the gap buy adding new thing on the research that already been cover most the staff. However predicting students performance dependence on a particular university and this research focused on a Witwatersrand university. My area of focus will to specialize on areas not covered by other publisher and the first time research of this nature is done at Wits.

For future work, the experiment can be extended with more distinctive attributes to get more accurate results, useful to improve the students learning outcomes. Also,experiments could be done using other data mining algorithms to get a broader approach,and more valuable and accurate outputs. Some different software may be utilized while at the same time various factors will be used.

References

[Adhatrao *et al.* 2013] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha, and Vipul Honrao. Predicting students' performance using id3 and c4. 5 classification algorithms. *arXiv preprint arXiv:1310.2071*, 2013.

[Aksenova *et al.* 2006] Svetlana S Aksenova, Du Zhang, and Meiliu Lu. Enrollment prediction through data mining. In *Information Reuse and Integration, 2006 IEEE International Conference on*, pages 510–515. IEEE, 2006.

[Al-Radaideh *et al.* 2006] Qasem A Al-Radaideh, Emad M Al-Shawakfa, and Mustafa I Al-Najjar. Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006.

[Bunkar *et al.* 2012] Kamal Bunkar, Umesh Kumar Singh, Bhupendra Pandya, and Rajesh Bunkar. Data mining: Prediction for performance improvement of graduate students using classification. In *Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on*, pages 1–5. IEEE, 2012.

[Dekker *et al.* 2009] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: A case study. In *Educational Data Mining 2009*, 2009.

[Hill *et al.* 1987] Thomas Hill, Nancy D Smith, and Millard F Mann. Role of efficacy expectations in predicting the decision to use advanced technologies: The case of computers. *Journal of applied psychology*, 72(2):307–313, 1987.

[Kabakchieva 2013] Dorina Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1):61–72, 2013.

[Kotsiantis 2012] Sotiris B Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.

[Kovacic 2012] Z Kovacic. Predicting student success by mining enrolment data. 2012.

[Luan 2002] Jing Luan. Data mining and its applications in higher education. *New directions for institutional research*, 2002(113):17–36, 2002.

[Lykourentzou *et al.* 2009a] Ioanna Lykourentzou, Ioannis Giannoukos, George Mpardis, Vassilis Nikolopoulos, and Vassili Loumos. Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology*, 60(2):372–380, 2009.

[Lykourentzou *et al.* 2009b] Ioanna Lykourentzou, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis, and Vassili Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3):950–965, 2009.

[Minaei-Bidgoli *et al.* 2003] Behrouz Minaei-Bidgoli, Deborah A Kashy, Gerd Kortemeyer, and William F Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in education, 2003. FIE 2003 33rd annual*, volume 1, pages T2A–13. IEEE, 2003.

[Nandeshwar and Chaudhari 2009] Ashutosh Nandeshwar and Subodh Chaudhari. Enrollment prediction models using data mining. *Retrieved January*, 10:2010, 2009.

[Nghe *et al.* 2007] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. A comparative analysis of techniques for predicting academic performance. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pages T2G–7. IEEE, 2007.

[Osmanbegović and Suljić 2012] Edin Osmanbegović and Mirza Suljić. Data mining approach for predicting student performance. *Economic Review*, 10(1), 2012.

[Romero and Ventura 2007] Cristobal Romero and Sebastian Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[Sacin *et al.* 2009] Cesar Vialardi Sacin, Javier Bravo Agapito, Leila Shafti, and Alvaro Ortigosa. Recommendation in higher education using data mining techniques. In *Educational Data Mining 2009*, 2009.

[Yadav and Pal 2012] Surjeet Kumar Yadav and Saurabh Pal. Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*, 2012.