UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

SCHOOL OF COMPUTER SCIENCE AND APPLIED MATHEMATICS

FACULTY OF SCIENCE

# Predicting the Completion of a Student's Science Degree based only on their First-year Marks :

Literature Review

Prince Ngema 754774

Supervised by:
Dr Ritesh Ajoodha and Dr Ashwini Jadhav

# 1 Introduction

Monitoring the progress of students' academic performance is an vital issue to higher learning institutions [Oyelade et al. 2010]. Studies show that the biggest churn rate occurs at the end of first year studies (29 % of first-year students drop out at the end of their first-year studies in South Africa). Over and above, only 30 % of the total first-time student graduate after a five-year period [Scott et al. 2007, as cited in [Ajoodha and Jadhav 2017]]. Higher learning institutions are in need of intervention programs that will improve student retention rates and graduation rates, these programs require previous knowledge of students' performance [Yadav et al. 2012].
The vast usage of computers and the internet has increased the availability of data that can be analyzed in order to predict students' performance. Using this data, researchers have attempted to predict students' performance in various higher learning institutions. They make use of various techniques like data mining, statistical analysis and machine learning to predict students' performance.

Many a time, admittance into a Wits University programme is a transformative experience for matric students, it gives them and their families hope for a lustrous future. Sadly, most students who are accepted into the university programme fail to complete their degree due poor academic performances. These students are left lamenting and drowning in debt. In this research, we attempt to predict the completion of a student's Science Degree based only on their first-year marks, that is , we attempt to calculate the probability of student successfully completing a Science degree given their first year marks, determine the value of taking combinations of subjects towards the completion of a student's degree and calculate alternative streamlines that might align the student with better options, so that they can reconsider their academic standing.

This Literature review provides a description of how the this problem of predicting students' performance was tackled by other studies. It gives an in depth description of the methods used and results obtained. In section 2 we look at the background and related work, we look at a machine learning technique that is used in predicting students' performance and we review previous work conducted on this topic. In the last section, section 3 , a conclusion is given.

# 2 Background and Related Work

In order to calculate the probability of student successfully completing a Science degree given their first year marks, we make use of a machine learning technique (naive Bayes' classifier). In section 2.1 a brief description of machine learning is given. In section 2.2 we look at the naive Bayes' classifier. In section 2.3 studies done on the field of students' performance are reviewed.

## 2.1 Machine Learning

In simpler terms, it is defined as the ability of a machine to learn from experience [Mitchell 1997]. In Oyelade et al. [2010] it is defined as the process of learning from examples or instances. For a two-class problem, it can be informally defined as: Given a target variable **y**, a classifier **Z** and a set of examples (instances) X for which y is defined over, train **L** on **X** to estimate **y** [Oyelade et al. 2010]. Given input data, a computer can find dependencies in the data that are too complicated for the human eye to form

[Pojon 2017]. Two prominent types of machine learning are supervised[1] and unsupervised[2] learning [Mitchell 1997]. Input data can mostly either be structured or unstructured [Mitchell 1997] . If data is structured, machine learning can be used to find dependencies in the data and make predictions . In this research, we have structured labelled data, hence we will make use of a supervised learning technique.

## 2.2 Naive Bayes classifier

This algorithm is the most pragmatic learning approach for most prediction problems [Pojon 2017]|. It is built on evaluating probabilities for hypotheses [Islam et al. 2007].The Bayesian model rivals other learning algorithms and in many prediction cases outclasses them [Alpaydin 2009]. The model using Bayes' theorem classifies instances to one or a number of independent classes using a probabilistic approach [Koller et al. 2009]. It is the easiest learning algorithm to implement [Pojon 2017].

This section is structured as follows. In section 2.2.1 we look at Bayes' theorem. In section 2.2.2 we explore how the naive Bayes' classifier is trained. In section 2.2.3 we look at classification, how the classifier makes predictions. In section 2.2.4 we look at feature engineering , how features are selected and created in the data set to improve the results. In section 2.2.5 we look at how we evaluate the performance of the classifier using the confusion matrix.

### 2.2.1 Bayes' Theorem

Suppose we have features $X = \{x_0, \dots x_n\}$ and target classes $Y = \{y_0, \dots y_n\}$, the goal of this classifier is to determine the probability of the features occurring in each class and return the most likely class. To achieve this, we use the Bayes' theorem that computes the probability of an event occurrence, based on the probabilities of other events that influence it.
Koller et al. [2009] asserts that given two different events $A$ and $B$, Bayes' Theorem states that:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{1}$$

where

- $P(A \mid B)$ is the posterior probability
- $P(B \mid A)$ is the likelihood of B given A
- $P(B)$ is the class probability
- $P(A)$ is the prior probability of A.

[1] the input data is labelled into two or more classes
[2] input data is not labelled

### 2.2.2 Training

For given data set with features $X = \{x_0, \dots x_n\}$ and target classes $Y = \{y_0, \dots y_n\}$ we make use equation 1 and we replace A with class $Y_i$ and $B$ with features $x_0$ through $x_n$ to train the model. Features are conditionally independent given the target class [Koller et al. 2009] This implies that

$$P(x_0, ..., x_n \mid Y_i) = P(x_0 \mid y_i) * P(x_1 \mid y_i) * ..... * P(x_n \mid y_i)$$

Therefore according to Koller et al. [2009], the final representation of the class probability is

$$P(y_i \mid x_0, ...x_n) \propto P(x_0, ...x_n \mid y_i) P(y_i) \propto P(y_i) \prod_{j=1}^{n} P(x_i \mid y_i) \tag{2}$$

### 2.2.3 Classification

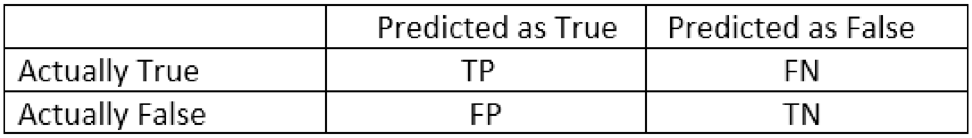
The problem of classification using the naive Bayes classifier can be informally stated as: Given datum $X = \{x_0, \dots x_n\}$ find a class in $Y = \{y_0, \dots y_k\}$ for $X$. Friedman et al. [1997] says classification is done by applying Bayes' rule to compute the probability of the features occurring in each class and then predicting the class with the highest posterior probability. This is called the maximum posterior hypothesis [Koller et al. 2009].

### 2.2.4 Feature Engineering

Feature engineering in machine learning is the process of selecting or creating features in a data set to improve the results of the prediction model [Domingos 2012]. Feature selection includes getting rid of unnecessary or redundant features. To remove these features (variables), we assess the relevance of the variable by fashioning a model to test the correlation of the variable with the dependent variable [Pojon 2017]. This can also be done manually by using domain knowledge [Domingos 2012]. Feature creation involves changing the variables and making new ones by combing multiple different variables [Pojon 2017].

### 2.2.5 Performance

To investigate the performance of a model, actual values must compared with predicted values [Pojon 2017] . The performance of the naive Bayes'model can be evaluated using a confusion matrix. It shows the type of classification errors [Oyelade et al. 2010]. Figure 1 serves as an example of a possible confusion matrix.



| | Predicted as True | Predicted as False |
|---|---|---|
| Actually True | TP | FN |
| Actually False | FP | TN |

Figure 1: Confusion Matrix

TP (True Positive) is the number of positive instances correctly classified, FP (False Positive) is the number of positive instances misclassified as negative, FN (False Negative) is the number of negative instances misclassified as positive and TN (True Negative) is the number of negative instances correctly classified [Oyelade et al. 2010]. Using the confusion matrix values, there are different types of performance evaluation criteria that can be used [Pojon 2017]. The first criteria Pojon [2017] discusses is termed Accuracy which is simply the ratio of correct predictions.

$$A = \frac{TN + TP}{TP + TN + FP + FN} \tag{3}$$

The second criteria in Pojon [2017] is called Precision and and Recall is given by

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

In th second criteria precision and recall are used together to make a better evaluation. A well performing predictive model must have a good conjunction of successful negative and successful positive predictions[Pojon 2017].This leads to the third criteria called the F-measure defined as Pojon [2017]

$$F = 2 * \frac{P * R}{P + R} \tag{6}$$

F-measure takes both precision and recall into account and outputs a single value.

### 2.3 Related work

Studies to predict students academic performance have been made over the years. Various techniques used to predict students' include data mining, statistical analysis and machine learning. In this section, we review some of the work done in the field of students' performance prediction.

A study to predict retention rate was conducted in Yadav et al. [2012]. Researchers from three different universities in India applied various machine models to predict retention rates. They compared these models and found that the ADT decision tree model provided most accurate results Yadav et al. [2012].

Another study conducted was to identify at risk-students[3] in Mathematical sciences using biographical data and enrollment observations to was conducted by researchers at the University of Witwatersrand, South Africa [Ajoodha and Jadhav 2017]. The basic methodology was to indicate influence of four biographical characteristics (i.e. gender, spoken home language, home province, and race description) on student aggregates, explore the trajectory of student performance over the period 2008 to 2017 respect to biographical characteristics and calculated the posterior probability of failing to complete the minimum requirements given various biographical profiles using Bayesian analysis. The results showcased at-risk biographical profiles with a Bayesian estimate that was greater than 0.7 for failing

[3] students at risk of not completing a degree in Mathematical Sciences.

to complete the requirements for a degree in the Mathematical Sciences.

However, predicting students' performance instead of identifying at-risk students is more related to the research topic. We have examples of such studies as well. One of these studies was conducted by Oyelade et al. [2010], they used 6 machine learning algorithms (Artificial neural networks, naive Bayes'classifier, Logistic regression ,KKN and SVM) to predict students' performance in distance learning. They also compared these learning algorithms and concluded that the Naive Bayes algorithm has more than satisfactory accuracy ,its overall sensitivity is extremely satisfactory and is the easiest algorithm to implement[Oyelade et al. 2010].

A similar study was conducted by Butcher and Muth [1985], they used American Collage Testing Program(ACT) test scores along with performance in high school and information regarding the students' programs to predict performance in an introductory Computer Science course and first semester collage grade point average. They took the statistical analysis approach and used the Statistical system (SAS) to perform all statistical analyses.

A recent study was done by Pojon [2017] , they investigated students' performance using machine learning. They used various machine learning techniques(Linear regression, Decision tress, naive Bayes' classifier) and compared their performance. Feature engineering techniques were used to improve prediction performance. They found that the naive Bayes' classifier performs better than the other the machine learning algorithms. Comparing the performance of these algorithms on engineered data and raw data, better performance results were obtained on engineered data.

## 3 Conclusion

The topic of this research is predicting the completion of a student's science Degree based only on their first-year marks. In this literature reviewed we attempted to provide background of the model needed for prediction and also review various studies on the field of predicting students' performance.

Starting by describing the naive Bayes' model, an in- depth description of the naive Bayes algorithm was given. The Bayes' theorem is the foundation of this algorithm. It was given in detail how this algorithm is trained and how it classifies. Model evaluation methods are also given. It was shown that feature engineering is important as it increases the performance of a model. According to most of the studies this algorithm is the easiest to implement. With the exception of Yadav et al. [2012], the naive Bayes' model was found to be the most effective algorthim having the highest accuracy in predicting students' performance. According to Koller et al. [2009], strong independence assumptions underlying this model decrease its diagnostic accuracy. If the accuracy of the model is not satisfactory, this model can be extended to Bayesian networks. Bayesian Networks have more have realistic independence assumptions [Koller et al. 2009].

The concept is similar in most of the studies studies reviewed. Given data, they apply different machine algorithms and build prediction models and perform comparisons between these models to see which one performs better. In Yadav et al. [2012], Pojon [2017] and Scott et al. [2007] the performance of these models is compared using accuracy, precision, and recall. Pojon [2017] goes further and introduces feature engineering so as to improve the performance of these models. Most of these models use biographical data to predict students performance, i.e Ajoodha and Jadhav [2017], uses biographical data and enrollment observations. In this research, only first-year marks will be used to predict students' performance.Similarly, uses first semester collage grade point average as one of the features to predict students' performance. The papers reviewed were very useful because they provide relavent information that will used in this research.

## References

Ritesh Ajoodha and Ashwini Jadhav. Identifying at-risk undergraduate students using biographical and enrollment observations for mathematical science degrees at a south african university. *Private Communication*, 1(1):1–21, nov 2017.

Ethem Alpaydin. *Introduction to machine learning.* MIT press, 2009.

DF Butcher and WA Muth. Predicting performance in an introductory computer science course. *Communications of the ACM*, 28(3):263–268, 1985.

Pedro M Domingos. A few useful things to know about machine learning. *Commun. acm*, 55(10): 78–87, 2012.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

Mohammed J Islam, QM Jonathan Wu, Majid Ahmadi, and Maher A Sid-Ahmed. Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, pages 1541–1546. IEEE, 2007.

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques.* 2009.

Tom M Mitchell. what is machine learning. *AI magazine*, 18(3):11, 1997.

OJ Oyelade, OO Oladipupo, and IC Obagbuwa. Application of k means clustering algorithm for prediction of students academic performance. *arXiv preprint arXiv:1002.2425*, 2010.

Murat Pojon. Using machine learning to predict student performance. Master's thesis, 2017.

Ian Scott, Nanette Yeld, and Jane Hendry. *Higher education monitor: A case for improving teaching and learning in South African higher education*, volume 6. Council on Higher Education Pretoria, 2007.

Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal. Mining education data to predict student's retention: a comparative study. *arXiv preprint arXiv:1203.2987*, 2012.