الجامعة البريطانية في دبي

The British University in Dubai

Performance Prediction Using Classification

التنبؤ بالأداء باستخدام التصنيف

by

GITA MOOLIYIL

Dissertation submitted in fulfilment of

the requirements for the degree of

MSc INFORMATICS

(**KNOWLEDGE AND DATA MANAGEMENT**)

at

The British University in Dubai

May 2019

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application

Gita M.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

The use of classification as a data mining approach for performance prediction has been studied by many eminent researchers. The objective of this study is to determine the best classification models for predicting At Risk status of students in their first semester of an undergraduate degree program. A comprehensive evaluation requires that multiple models with different algorithms were analyzed using key performance measures. Principal component analysis and feature selection by weights using information gain ratio, Gini index, correlation and PCA is used to determine the relevant predictors of the datasets used. This study also addresses gaps in the current available literature on performance prediction, such as data imbalance and the use of Ensemble models. Sampling and weighting techniques were included using Rapid Miner® operators for SMOTE, stratified, bootstrap sampling and weighting. Ensemble models using bagging, boosting and the vote operator in addition to Gradient Boosted trees and Random Forest were compared to the individual classifiers to measure model efficiency. The best models were then used to ascertain how early at-risk prediction can be employed using data on student performance in the course assessments. The results show that Ensemble models and the use of sampling and weighting clearly improves model performance. The early risk prediction as expected is most accurate with all the coursework and final grades in a semester. Interestingly the variance in the performance measure values are not very significant for some of the models and it can be concluded that early risk prediction can happen earlier in

the semester when intervention and associated benefits of improving student performance is more probable.

الملخص

العديد من البحثين قاموا بدراسة استخدام التصنيف كنهج لاستخراج البيانات من أجل التنبؤ بالأداء. الهدف من هذه الدراسة هو تحديد أفضل نماذج التصنيف للتنبؤ بحاله المخاطر لدي الطلاب في الفصل الدراسي الأول من برنامج درجه البكالوريوس. التقييم الشامل للطلاب يحتاج تحليل نماذج متعددة ذات خوارزميات مختلفه باستخدام مقاييس الأداء الرئيسية. يتم استخدام تحليل المكونات الرئيسية (PCA)، واختيار المعالم بواسطة الأوزان باستخدام نسبه كسب المعلومات، ومؤشر جيني لتحديد التنبؤات ذات الصلة بمجموعات البيانات المستخدمة. وتتناول هذه الدراسة أيضا الثغرات في الأدبيات المتاحة حاليا بشان التنبؤ بالأداء ، مثل عدم التوازن في البيانات واستخدام نماذج الفرقة. و تشمل ايضا تقنيات أخذ العينات والترجيح باستخدام مشغلي SMOTE و Rapid Miner®، الطبقية، وأخذ العينات التمهيد والترجيح. نماذج الفرقة باستخدام التعبئة ، وتعزيز والمشغل التصويت بالاضافه إلى التدرج عززت الأشجار والغابات العشوائية تم مقارنتها لمصنفات الفردية لقياس كفاءه النموذج. ثم استخدمت أفضل النماذج للتاكد من الكيفية التي يمكن بها استخدام التنبؤ المبكر بالمخاطر باستعمال بيانات عن أداء الطلاب في تقييمات المقرر الدراسي. وتبين النتائج ان نماذج الفرقة واستخدام أخذ العينات والترجيح يحسن بوضوح أداء النموذج. التنبؤ بالمخاطر المبكرة كما هو متوقع هو الأكثر دقه مع جميع المقررات الدراسية والدرجات النهائية في الفصل الدراسي. ومن المثير للاهتمام ان التباين في قيم قياس الأداء ليست كبيره جدا بالنسبة لبعض النماذج ، ويمكن استنتاج ان التنبؤ بالمخاطر المبكرة يمكن ان يحدث في وقت سابق من الفصل الدراسي عند التدخل والفوائد المصاحبة لتحسين الطالب الأداء هو أكثر احتمالا.

Acknowledgements

Doing this dissertation although challenging has been huge learning experience for me both in terms of my research skills and data mining knowledge. It has given me the exposure to enhance my abilities in data mining and research and kindled in me the desire to do more research and enhance my knowledge in data sciences.

I sincerely thank my teacher and supervisor Prof Khaled Shaalan for giving me the support and guidance and to Dr Sherief Abdallah for being my teacher and guide in the Data Mining course of the MSc Informatics program.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

“The UAE National Strategy for Higher Education 2030 emphasizes the need to provide students with the technical and practical skills to be productive elements that can push the wheel of economy in the public and private sectors” (National Strategy for Higher Education 2030, no date).

Aligning with this strategy makes it imperative for higher education universities in this region to address student attrition. Student performance and graduation rates in higher education universities will be a key factor in ensuring success in achieving this strategy. Higher education universities should therefore employ effective techniques in ensuring that a high percentage of their students graduate and become key partners in driving the purpose and vision of the UAE. Predicting students At Risk is a fundamental requirement which will help in lowering attrition, motivating learners and ensuring better student performance. Employing intervention methods for students who are identified ‘At Risk’ can promote learning and student involvement which aligns with Tinto’s retention model that is most widely discussed in higher education literature. “Learning has always been the key to student retention”(Tinto, 2007). The objective of this research study is to predict if students in their first year of University in a local institution in the UAE are at risk of failing or dropping out from the course or program. The current practice at the University is to identify students At Risk midway through a 16-week semester based on course assessment performance and a high percentage of absences. As the efficacy of this approach needs to be improved the purpose of this study is to determine whether a suitable data mining approach can be used to accurately predict At Risk for students who are entering the program based on pre-college data and other relevant academic information. Various data mining algorithms will be used to build the performance prediction models. The best models will then be used with the course assessment data to investigate the reliability of early prediction of At Risk based on course assessments that students have completed. This two-fold approach of identifying students At-Risk at the start of a semester and then using course performance data to monitor their At-Risk status will allow for intervention at an earlier stage which could lead to more students being successful at the end of the first semester and higher rates for course completion.

The importance of using education data mining as an approach has been highlighted in large number of studies. Researchers who have contributed to work done on predicting students at risk stress on the importance of early identification of students at risk as early as Semester 1 (Agnihotri and Ott, 2014). Features such as college admission test scores and high school grade point average are important predictors of At Risk.(Raju and Schumacker, 2015) Demographic attributes such as gender, race and financial need are known to be good predictors but do not necessarily contribute to instructor intervention. (Baker *et al.*, 2015) Due to privacy and other constraints the datasets used in this study will only consider course grades and other relevant information. No personal or social information of the students is included in this study.

The objective of this research project will be to study the current literature done both within the UAE and outside the UAE on performance prediction and use it as a basis to design a data mining solution to address the business problem of identifying students at risk. Apart from adopting best practices within the literature reviewed, the dissertation will also address gaps in the current literature on prediction performance and attempt to include this if relevant to the objectives of this study.

The CRISP-DM framework proposed by Kotu and Deshpande (2014) will be adopted as the structured process for the data mining solution to the business problem of student performance prediction.

Research Questions

The following questions have been formulated to direct the research on this topic and satisfy the objectives:

1. What are the most effective and efficient data mining algorithms for predicting students at risk using the given dataset?

    1a. How do ensemble models perform in comparison to the models using individual classifiers?

    1b. Does the use of sampling and weighting techniques improve model performance?

2. What are the highly relevant predictors in the given dataset?
3. How early can we accurately predict student ‘At Risk’ status in terms of the course work done in Semester 1 of the program?

Chapter 2 Literature Review

A literature review of current work is conducted in this study using resources such as Scopus, ProQuest, Elsevier, IEEE and others. Books and articles on data mining as well as the Rapid Miner documentation were also used as resources.

Supervised data mining techniques will be used in this research project to evaluate model performance in predicting students at risk in their first year of University. Classification is a data mining technique that is popularly used to predict student performance. (Shahiri, Husain and Rashid, 2015). Several studies in the area of student performance prediction have recommended and used Decision tree, Naive Bayes, Neural Nets, SVM and Logistic Regression for binominal classification.

Classification methods employed is based on the literature review of similar studies that focus on predictive models used for labelling students who are At Risk of failing or dropout from a course or semester. A summary of the literature review is provided below. The literature review has sub sections that refer to the work done by current research studies in the specific areas. These sub sections focus on the main topics of this study.

Educational Data Mining (EDM)

Educational data mining is concerned with developing methods to deal with the unique type of data that is peculiar to educational settings. It helps to better understand students and the environment that they learn in.(Baker and Yacef, 2009). This sentiment about EDM has been echoed in the work done by Romero and Ventura, (2010) who also state that the process of EDM which converts raw data from education systems into something meaningful that has a significant impact on education practice and research. Studies done by researchers agree that the popular data mining tasks are clustering, classification and association analysis while decision tree, Naïve Bayes and Neural Net are the common algorithms used (Romero and Ventura, 2010; Dutt, Ismail and Herawan, 2017). A majority of studies have used multiple algorithms with classification to study educational data in their environment in order to predict performance and identify students At Risk. (Jayaprakash *et al.*, 2014; Lakkaraju *et al.*, 2015; Raju and Schumacker, 2015; Costa *et al.*, 2017). Other research also endorses this approach in their study when they discuss how student modeling can help determine factors that are predictive of At Risk or non-completion of college courses or college itself (Baker and Yacef, 2009)

Framework

CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) is one of the most popular and widely adopted framework for data mining solutions. (Kotu and Deshpande, 2014). CRISP-DM is defined as a process model for data mining solutions that is divided into six phases that fit into a natural loose order to accommodate real world implementations that could be different and unique to the business situation (Putler and Krider, 2015). Others justify the use of CRIPS-DM as a research approach because it is nonproprietary and application neutral and hence relevant for all data mining projects (Kabakchieva, 2013). The six stages of this model include business understanding, data understanding, data preparation, modeling, evaluation and deployment. The internal feedback loops between phases allows for a nonlinear approach that can achieve results that are consistent and reliable. Some researchers have made use of this framework in their study on prediction of employment status of fresh graduates using supervised and unsupervised learning. (Azziaty Binti Abdul Rahman, Lam Tan and Kim Lim, 2016)

The CRISP-DM framework provides a structured approach to determining the solution to a data mining research project. It will be used in this study to detail the processes covered in building a model to predict performance and classify students at risk.

Predictive Modeling

This section starts with a review of research work done on performance prediction using classification within the UAE followed by a review of other literature in this domain.

UAE research

Recent papers in the area of student performance prediction for Universities in the UAE studies were reviewed as the student population being considered in this study are also from the UAE. (M., F. and A., 2018) discuss the use of logistic regression to predict academic performance in the final exam based on the assessments done during the year in a University level programming course in the UAE. Logistic regression is used to evaluate the impact of the course work assessments done on the final score using t-stat and p values. No performance measures such as accuracy or other values have been considered in this study. The use of a single classifier also limits the ability to compare performance accuracy amongst different classifiers in the given dataset. The effect of personal and social factors on the previous performance of over 250 students at different colleges in in Ajman University of Science and

Technology (AUST), Ajman, UAE has been studied in the research done by (Abu Saa, 2016) . Student performance is based on GPA in this case. Classification is the data mining method used in this study by comparing the performance of multiple classifiers including four decision tree algorithms C4.5 decision tree, ID3 decision tree, CART decision Tree, and CHAID and the Naïve Bayes algorithm. Naïve Bayes was also used to determine interesting and non-interesting attributes based on their probability values. A multi class label corresponding to the GPA value was used as the target variable using nominal values such as Excellent, Very Good, Good and Pass. The concept of a default model is used as a baseline for measuring model performance. Accuracy, recall and precision were used to compare model performance and CART was the best performing decision tree with an accuracy of 40% which was relatively higher than the default model. Imbalance was only considered in the second study but apart from the mention of using random sampling no details were provided on the sampling operators used in Rapid Miner® or Weka®

Research done outside the UAE

Supervised machine learning techniques such as Decision tree (CART) and Random Forest are used to determine the course work assessments to predict student performance in a Thermodynamics course for a dataset of 36 examples by (Akangah et al., 2018). Accuracy, classification error and kappa scores were compared for both models, and the tree structure identifies the attributes that influence prediction. Kappa scores for the CART decision tree model with cross validation and the Random Forest models varied between 0.4 and 0.78 showing moderate to substantial agreement. Models are better at predicting the Pass rate rather than the Fail status. Since the models used are both variation of the decision tree algorithm it does not clarify whether classification with other models such as Naïve Bayes, SVM or Neural Net would result in higher prediction accuracy

A first year Student At-Risk Model (STAR) is designed to identify key risk factors for freshmen, enabling administrators to employ early intervention to prevent or decrease student dropout (Agnihotri and Ott, 2014). The At-Risk model was built using four different models namely Logistic regression, Naïve Bayes, Neural Net and Decision Tree using Year 1 data from NYIT. Sample size is about 1453 students. Multiple versions of the models were run with parameter variations and the models with the highest recall were used in the Ensemble Model. The ensemble model uses all of the student data as well as the output of the four models combined as input. If recall values were similar, then the higher precision value is used as the filter. Recall values were about 70% while precision values were about 50. Highest recall achieved was about 75%. The discussion on this study seemed to suggest the proportion of at-risk students were typically less than the students who were not at risk. No attempt has been made to consider data imbalance and use the necessary techniques to counter this. Studies state that EDM techniques such as Decision Tree, Naïve Bayes SVM and Neural Net are effective in early prediction of students who are likely to fail based on the f-measure values of model performance. (Costa *et al.*, 2017). The other conclusion in this study is that data preprocessing and parameter fine tuning do not necessarily improve model performance. As this study only uses one performance measure the relevance of this claim could be argued. There is no mention of the data characteristics in terms of imbalance and this is another factor that could possibly invalidate their model performance values. Another research study proposes a machine learning framework to identify and prioritize intervention methods for student data from two US districts who are at risk of not graduating in time (Lakkaraju *et al.*, 2015) . The focus of the evaluation process and results is not to merely achieve high performance measure scores but to the cater to the needs of the instructors who want a comprehensible framework that they can use. Accuracy, recall, precision and AUC are used to evaluate performance using traditional metrics. Classification models used include Random Forests, Logical Regression, Adaboost, SVM and Decision Tree. A risk ranking scheme is used to group students at different risk levels to enable schools to plan intervention based on resource availability. An empirical risk curve is used to analyze risk scores. Quality is assessed based on an empirical curve that is monotonically non decreasing. According to the researchers Random Forest consistently produces good results and hence is recommended for ranking students from both districts. Precision and recall curves are developed for different values of K where K denotes the number of students that could possibly have interventions based on school resources. Using a K of 5% results in Random Forest outperforming all other models for both school districts. Providing the ability to use a K threshold for these models is a novel approach and empowers the instructors to resource reliable intervention. Although the objective of the study caters to high school students the model could be valid for other educational settings too. Using precision recall curves to evaluate performance has been seconded by other studies but including measures such as geometric mean or kappa scores could also add value by using multiple measures and this would still meet the criteria of a simple framework that instructors can comprehend. Asif et al., 2017 proposes a detailed study to answer three different research questions. The questions are whether accurate prediction for at risk students is possible in the early stages of a program using student grades only, can they identify courses that influence student performance, and whether student progression can be related with the indicator course. Multiple classification models are used to determine the most accurate model performance. Classifiers are then chosen to predict the courses that influence student performance based more on interpretability than accuracy. Clustering is then used to group students that have a similar progression during the four-year course. The information from the clustering algorithm provides input to effective interventions for at risk students. Various classifiers were used with decision trees and random forests using a varied combination of parameter settings. Decision trees were used to implement feature setting based on the attributes that were selected as internal nodes or leaf nodes. Accuracy and kappa scores were used to compare the performance measures for these classifiers. Naïve Bayes was the best performing model and achieved a high accuracy of over 80%. (Asif *et al.*, 2017). The novel method of feature selection using decision trees as mentioned in the study had better accuracy than other feature selection methods. There is a danger here that if the splitting criteria of a decision tree is inaccurate then important attributes could be ignored. Perhaps an option is to compare the decision tree nodes with the feature selection options of attributes and then make a judgement using this and domain understanding of the attribute set. There is a mention of data imbalance in the study, but no techniques or measures have been applied to counter the effects. This study has focused on predicting student dropout in higher education is using logistic regression. (Kang and Wang, 2018) The predictor variables are gender, ethnicity, time status, classification, age, earned hours, and overall GPA. Performance measures of accuracy, precision, specificity and recall are used to evaluate the models. The researchers observe that there is a significant increase in recall when the data is balanced i.e. random samples of negative instances are chosen to match the number of positive instances where dropout = yes. Other classifiers including k nearest neighbor (k-NN), Decision Tree, Naïve Bayes, Support Vector Machines (SVM), and Random Forest are used to predict the dropout for comparison purposes. F measures are recorded to study the precision and recall rates. An overall relatively high overall accuracy rate of 81.8% for unbalanced data and recall rate of 75.9% for balanced data respectively is achieved. Since the researchers have observed better performance with manually balancing the dataset, techniques such as sampling or weighting could have been employed to possibly enhance performance scores. Coniin et al. (2017), in their study assess the performance of neural networks in comparison to other classifiers in predicting student performance using LMS data of about 4601 students. Another aspect of their research is to determine how omitting the course ID from the attribute sets affect the results and finally to determine if using individual course data for the study versus all of the data from different courses affect the performance of neural networks. The other classifiers used in this study are k-Nearest Neighbors, Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree and Random Forests. Accuracy and recall were used to measure performance. The Neural Network outperforms all other classifiers when it comes to accuracy with a score of 66.1% followed by Logistic Regression which has an accuracy score that is 3.5% lower than the Neural Network, at 62.4%. It was also determined that sample sizes had no effect on the accuracy of the model's performance. Removing the course ID as an attribute did not affect the accuracy of Neural network but it suffered a lower recall than other algorithms used. Variance in the performance measure values without the course ID predictor was minimal. Conducting the modeling for the dataset for individual course had a varied performance with an increase in accuracy for some of the courses while it decreased in others (Conijn and Zaanen, 2017). There is no discussion on whether there is any data imbalance as this could also have affected the efficiency of the scores obtained.

Jayaprakash et al. (2014) work on performance prediction is one of the few studies in this literature review with high number of examples, over 9000 examples for the training set and 5212 examples for the testing dataset (Jayaprakash et al., 2014) . Binominal classification was compared using logistic regression, support vector machines using sequential minimal optimization (SVM/SMO), J48 decision trees, and Naïve Bayes classifiers. Measures used were the values of accuracy, recall, false positives, false negatives, true positives and true negatives. Predictive performance was analyzed on both the imbalanced data and multiple samples of the balanced data which was sampled at varying rates of 25%, 50%, 75% and 100% of the training data. The values recorded for recall in three of the classifiers is over 80% while maintaining low false positive values at about 17%. The probability of At Risk was lower in full time students when compared to part time students. High values of course grade that contributes to the final grade, online Sakai sessions and cumulative GPA reduced the probability of being at risk. Logistic regression was used to compute these coefficients. Although it is suggested that the samples were balanced use of sampling or weighting to address the issue of imbalanced data could have been investigated to evaluate impact on model performance. Adejo et al. (2018) studies classification of a student dataset with 143 example sets and uses principal component analysis to select the significant student related attribute. Modeling is performed using three base classifiers and ensemble classifiers to compare efficiency of performance across these models. Stacking is used as the ensemble learning models where the output from multiple base classifiers serves as the input to a meta classifier for the final prediction. Performance measures included accuracy, the precision and recall, the F-measure, classification error and the root mean squared error (RMSE) of the three classifiers. Higher values for f-measure, accuracy, precision and recall were observed for the ensemble model which also recorded low values for RMSE and the classification error (Adejo and Connolly, 2018). A lot of emphasis has been placed on using the proper set of attributes for performance efficiency but there is no mention of whether the data is imbalanced. Lehr et al., 2016, built three different predictive models on which six different data mining algorithms were applied. The models were as follows:

Pre college model which is based on pre-college datasets to predict a graduation rate for newly admitted students, first term model that includes first semester GPA and the difference between high school and first semester GPA, first term class hours and first term math tier, the third model is with all the data for the first two models in addition to cumulative GPA and cumulative GPA difference with high school GPA. The six algorithms were used to train and test models: Logistic Regression, Naïve Bayes, K Nearest Neighborhood (KNN), Random Forest, Multilayer Perceptron (MLP), and Decision Tree. Feature selection is used to determine the attributes that affect prediction accuracy for classifying students at risk. Best predictors of attrition are unweighted high school GPA, SAT/ACT scores, first year GPA and EFC (expected family contribution). The study also determined that early ‘at risk’ prediction can be achieved by the end of the first semester with 70% accuracy and first term GPA drop is highly indicative of the student being at risk of failing (Lehr et al., 2016). Although a detailed study of attribute impact has been conducted in this study data features that affect model performance such as data imbalance have not been discussed.

Class imbalance is a crucial factor to consider when using data mining for educational datasets as the classes are usually imbalanced. This affects model performance and hence the measures used to evaluate the model also differs. A summary of the few studies that address imbalance is provided below.

Class Imbalance

Thammasiri et al., (2014) address the issue of machine learning being unable to predict the less representative class in a dataset. This issue of class imbalance is addressed in this study when predicting performance for a student dataset that contained seven years of freshmen data from 2005-2011. The dataset consisted of 34 variables and 21,654 examples, of which 17,050 were positive/retained (78.7%) and 4,604 were negative/dropped-out (21.3%). Four popular classification algorithms were used in this study namely artificial neural networks, support vector machines, decision trees and logistic regression with three types of sampling random oversampling, random under sampling and SMOTE Synthetic minority oversampling technique to build prediction models. Performance measures used for this study included accuracy, sensitivity, specificity, precision+, precision-, F-measure, Correlation coefficient and Geometric mean. SVM with SMOTE balancing technique was reported as the best performer across the various measures employed in this study (Thammasiri et al., 2014). Zou et al., (2016) argues that although AUC is considered a reliable performance measure for imbalanced data, a high value of AUC may not accurately reflect the classification performance. This study recommends the use of F-score also known as F-measure in addition to AUC to ensure that model performance is correctly measured. The study proposes that examining different probability thresholds for the testing set will result in better accuracy for model performance measures. Random Forest was the classification technique used in this study with a medical dataset that is commonly used to evaluate performance of a range of homology detection methods (Zou et al., 2016). In an alternative approach Rashu et al., (2003) use only the performance measure of accuracy to compare how Decision Tree, Naïve Bayes and Neural net models perform on student final grade prediction with random oversampling, random under sampling and SMOTE employed to balance the data. Neural Net with SMOTE achieved the best performance accuracy comparatively of over 73%. Their research establish that all the three models perform with better accuracy values when sampling is enabled (Rashu, Haq and Rahman, 2003). Jeni et al., (2013) emphasize the importance of using the correct metrics with imbalanced data when classification models are applied to facial action unit recognition. Their findings are contrary to what is proposed by the above-mentioned studies. According to their research and tests AUC is the only value that is not affected by data imbalance although there is a mention of poor performance being masked by ROC curves (Jeni, Cohn and De La Torre, 2013).

Ensemble Models

Ensemble models combine multiple base classifiers and usually achieve a better performance than the individual base models. Bagging, boosting and stacking are the three main approaches to building ensemble classifiers.(Adejo and Connolly, 2018) Random Forest and Gradient Boosted Trees are also ensemble classifiers that use bagging and boosting techniques for ensemble modeling. In the study conducted by (Adejo and Connolly, 2018) the ensemble classifier using stacking had a performance that surpassed the base models in predicting academic performance of University level students. The classification techniques used in this study were Decision Tree, Artificial Neural Networks and Support Vector Machines (SVM). Performance measures used were Accuracy, F-measure, Precision, Recall and Classification error. (Kotsiantis, Patriarcheas and Xenos, 2010) describe the importance of an online ensemble classifier which is able to adapt when new training instances are seen the unknown knowledge is comprehended an incorporated immediately which then becomes useful for where data is being generated continuously as in a distance learning environment. This incremental learning ability of the model enables machine learning to be useful in a real time environment. The proposed online learning ensemble algorithm is built with Naïve Bayes, Winnow and 1-NN using a simple majority voting system. According to the researchers the tests revealed that this ensemble model performed much better than the individual base models.

Kumari et al., (2018) studied the effect including the behavioral features of students in an e-learning or web-based environment on student performance with classification algorithms such as k-NN, ID3, SVM and Naïve Bayes. Accuracy recall and f-measure recorded better marginally better values when behavioral attributes were included in the dataset while the precision value was slightly lower. The researchers also used Ensemble models such as bagging, boosting and voting with the four base classifiers. The voting method in ensemble models secured the best accuracy of about 89% (Kumari, Jain and Pamula, 2018). The use of sampling with and without replacement is referred to in this study but the issue of data imbalance has not been addressed.

Other studies such as the work done by Adejo et al., (2018) states that challenges in identifying students at risk due to low performance and accuracy is attributed to insufficient variable usage and the use of single base classifiers in predictive modeling (Adejo and Connolly, 2018). Furthermore, other researchers also reinforce the benefits of using ensemble models to achieve significantly higher values in performance and thereby increased accuracy . (Agnihotri and Ott, 2014; Amrieh, Hamtini and Aljarah, 2016). Stapel et al., (2016), eloquently describes why ensemble models are successful in finding a prediction where the single classifier fails. This is possible because ensemble models blend prediction from multiple classifiers to achieve better accuracy and also benefit from an improved generalization due to the use of different specialized classifiers (Stapel, Zheng and Pinkwart, 2016).

As there are fewer studies comparatively that use ensemble models for performance prediction one of the goals of this study will be to evaluate the use of ensemble model accuracy in predicting student risk in comparison to the other single base classifier models.

Summary

The literature on predicting student performance clearly indicates that most studies have used multiple classification algorithms for class prediction. Decision tree algorithms or its variants such as ID3, C4.5, J48 are commonly used in the tree category of algorithms. Logistic regression has also been used by multiple studies as a modeling technique. The use of SVM and Naïve Bayes is seen to be a popular choice in most of the literature reviewed. Neural Networks also referred to as ANN has been used as a classification algorithm for performance prediction in a few of the studies. There is no clarity on the best classification algorithm or group of algorithms that are recommended for student performance prediction. Since the dataset, features or attributes, sample size and the performance measures are unique to each study it does not allow a single model to be selected as the best option for predicting student performance.

The use of performance measures in the reviewed studies does not indicate consistency amongst researchers. Although accuracy, precision and recall are used by many of the studies some of the researchers caution against the use of accuracy from the confusion matrix as a single measure of model performance. This is more evident in the case of data imbalance which is not addressed in all studies. When the major and minor class in binominal classification is not balanced it is recommended that performance measures include the use of AUC, accuracy, f-measure and geometric mean.(He and Garcia, 2009; Thammasiri et al., 2014; Zou et al., 2016)

Most of the studies that were reviewed either did not consider data imbalance or the use techniques such as sampling or weighting to deal with the minority class in predicting performance. The significance of recognizing that most datasets are imbalanced is clearly outlined in the studies on prediction performance that focus on this issue. Hence, when employing modeling techniques for educational datasets the performance measurement parameters should be chosen carefully to ensure that multiple values such as AUC from ROC curves, f-measure and geometric mean are evaluated along with accuracy precision, recall and kappa scores. This will reduce the inconsistencies of using a single measure such as accuracy to evaluate model performance.

The benefits of using ensemble models which can harness the power of multiple base classifiers and effective improve prediction performance is clearly highlighted in the literature reviewed. As the few studies that have used Ensemble models effectively vary in terms of their approach and operators used, this research will essentially compare all ensemble models such as

Random Forest, Gradient Boosted trees, Bagging, Boosting and using the Vote operator to build ensemble models with groups of two and three classifiers.

Research rationale

Reviewing the large body of research work done in the field of student prediction performance in the field of educational data mining it is clearly established that using a supervised data mining approach such as classification is predominant in most of the studies except for a couple of studies that also included clustering as an approach.

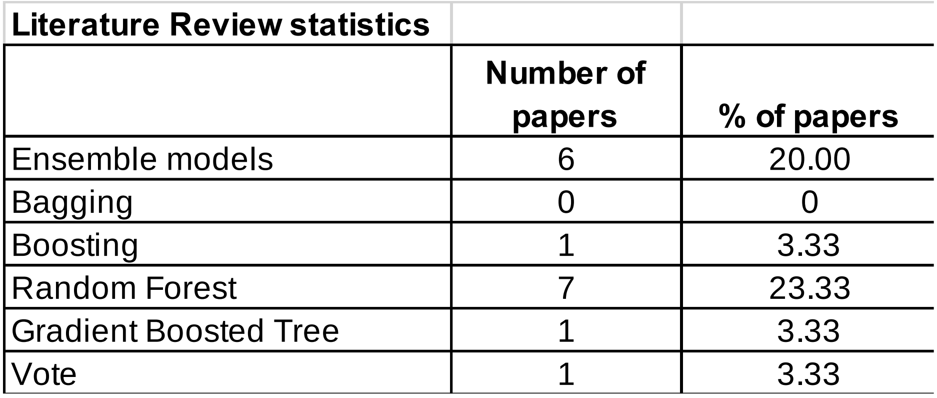
A list of the papers reviewed as part of studying the current available literature on prediction performance is provided in Appendix 1. The summary table of the number of papers that use classification and the type of algorithms used is listed below

Literature Review statistics		
	Number of papers	**% of papers**
Data mining approach and algorithms		
Classification	30	100.00
Decision Tree	22	73.33
Naïve Bayes	19	63.33
Neural Net	9	30.00
Logistic Regression	15	50.00
SVM	10	33.33

Table 2—1 Literature Review Statistics Algorithms

It is clearly established from this review that classification is the data mining approach used by all the studies and almost all of the 30 papers reviewed for the literature review use multiple classifiers to evaluate performance on predictive modeling. 73% of the papers reviewed use Decision tree followed by

Naïve Bayes which was used by 63% of the papers. Logistic regression, SVM and Neural Net have a lower percentage of occurrence in the papers reviewed. Ensemble models were used in very few papers and the techniques employed was more of Random Forest, and only a few studies had used Boosting, Vote and Gradient Boosted Trees. The summary of work done in this area is listed in the table below.



Literature Review statistics		
	Number of papers	**% of papers**
Ensemble models	6	20.00
Bagging	0	0
Boosting	1	3.33
Random Forest	7	23.33
Gradient Boosted Tree	1	3.33
Vote	1	3.33

Table 2—2 Literature Review Statistics Ensemble Models

Class imbalance and measures to counter this such as random sampling and SMOTE was used in 4 papers while weighting was mentioned in one of the papers but not implemented.

The major objective of this dissertation is to investigate the best classification models using a data mining approach to predict student performance based on pre-college performance data and use these models to measure the degree of accuracy that can be achieved in early prediction of At Risk using the course assessment grades in Semester 1. The aim of predicting performance is to evaluate whether the students is at risk of failure and eventually being dismissed from the program.

The central question of this research is how effective data mining algorithms in are predicting student performance. This leads to further investigation of current research to determine the data mining approach and algorithms that can be used. Data imbalance which is prevalent in educational datasets will be considered and the impact of using sampling and weighting techniques will be studied, in addition to class imbalance, the use of ensemble models is also rare, and this will be included to ascertain the best model performance for this dataset. These findings will result in specific data mining algorithm models being selected based on multiple performance measures. The best models will then be used on dataset 2 to identify how accurate early prediction of at risk is using multiple models based on assessments completed at different stages of the semester.

The conclusion from the literature review summary is to use classification as the data mining approach for predictive modeling. The use of multiple classifiers to evaluate and compare performance to decide on the most appropriate classification algorithms is also concluded. As Decision tree algorithms are easy to comprehend and resulting the tree structure clearly gives adequate information on the classification process it is popularly used even though the accuracy of the performance measure values is not always the best. The other algorithms do not provide much information on the process of classification and is more of a black box approach. Ensemble models such as Gradient Boosted trees and Random Forest also provide tree description identifying which are the relevant attributes used for data splitting.

The approach in this research will be to use algorithms that are used by the majority of researchers such as Decision Tree, Naïve Bayes and Logistic Regression and the algorithms used in a few studies namely Neural Net and SVM. As each classifier has its own special method for classification it will be prudent to include multiple base classifiers and then judge performance. It is clear that Ensemble models are not used in many studies. This is a gap that will be addressed in the study to investigate how these models perform in predicting At Risk for students. Researchers in this area have not dealt with class imbalance and studies its effects on model performance. As imbalanced classes are common to datasets used in predictive modeling this is another gap that will be addressed and studied in this research project.

Chapter 3 Methodology

Abbreviations and Acronyms used

Abbreviation/Acronym	**Description**
BS	Bootstrap Sampling
CRISP-DM	Cross Industry Standard Process for Data Mining
DT	Decision Tree
GBT	Gradient Boosted Trees
ID3	(Iterative Dichotomiser 3)
LR	Logistic Regression
NB	Naïve Bayes
NN	Neural Net
RF	Random Forest
SMOTE	Synthetic Minority Oversampling Technique
SS	Stratified Sampling
SVM	Support Vector Machines
W	Weighting

Table 3—1 Abbreviations and Acronyms

CRISP-DM Framework

Cross Industry Standard Process for Data Mining CRISP-DM has been used as a framework in some of the research work reviewed in educational data mining. The diagram below shows the different phases of this framework. (Kotu and Deshpande, 2014)

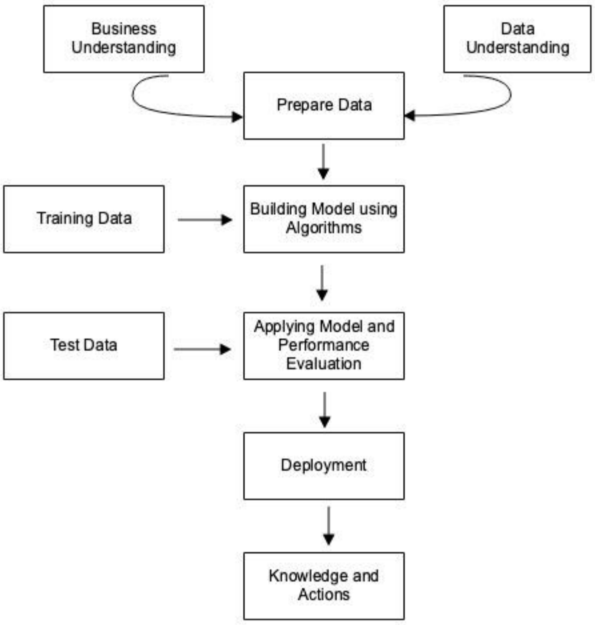


Figure 3-1 CRISP- DM Framework (Kotu and Deshpande, 2014)

The framework is implemented in this study and the description of the different stages of data mining study is described below. The implementation is adapted from Rahman et al.(2016) who made use of this framework in their study on prediction of employment status of fresh graduates using supervised and unsupervised learning and the information provided by Kotu et al., (2014). (Kotu and Deshpande, 2014; Azziaty Binti Abdul Rahman, Lam Tan and Kim Lim, 2016).

Business understanding

Business goals need to be considered in defining the problem that we are trying to solve using data mining. In this study the business requirement is to ensure that all students are supported to succeed in their higher graduation journey and in this context, we need to determine students who are ‘At Risk’ of failing. Predicting ‘At Risk’ earlier in the semester is the goal so as to ensure that there is enough time for intervention and helping the student succeed.

Data understanding

In order to satisfy the business objectives, it is important to have a clear understanding of the context, the subject which is in this case the student performance, assessment and grading process and the business process that generates the data. Student data that was used in this study only included relevant attributes to identify ‘At Risk’. Some of the researchers recommend using attributes that detail persona and social information for better prediction performance. In compliance with privacy and other guidelines no data that described social of personal features of the student were used. The binominal classification was done on the ‘Academic Standing’ feature that was created from cumulative GPA scores for Dataset 1. Students that were below a GPA of 2.0 were considered as ‘At Risk’ while those with a higher GPA were labelled as ‘Good Standing’. In dataset 2 the same theory was applied to create a feature called ‘GPA rating’. Here the same criteria are used for consistency and students at GPA below 2 are identified as ‘At Risk’. Campus details and gender were changed to numeric values by using a range of numbers to represent the campus value. Gender is replaced by 1 for Male and 2 for Female. Identifiers such as ID and username were also omitted from the modelling dataset. ID was used to join information from the two separates files that were used for Datasets 1 and 2. Academic Standing is the label for Dataset 1 while GPA-rating is the label for Dataset 2.

Data Preparation

The datasets were cleaned to remove all personal data, social information if any and other irrelevant attributes.

Data cleaning

Data preprocessing is one of the arduous and time-consuming processes of data mining. As most of the data mining algorithms require data in a structured format with attributes in columns and records in rows and business data is usually not structured in this format, data needs to be converted according to the algorithm requirements. (Kotu and Deshpande, 2014).

The student data was downloaded and then converted to Microsoft Excel format. Data was stored in separate files and these were combined using MS Excel functions such as VLOOKUP. The total number of attributes in the combined file was 100 this was reduced to 14 attributes for Dataset 1. Course grade information was not included in the model attributes as the aim was to find the best prediction model for identifying students At Risk at the start of the semester. The total number of records in Dataset 2 was combined from three different files that contained course grades and other relevant information such as cumulative GPA, high school average and IELTS scores with a total of 24 attributes in 435 records.

Data Quality

The accuracy of a model performance is dependent on the quality of the data provided to the model. Data quality requires cleansing to remove duplicate records, data entry errors, standardizing values, addressing outliers and transformation to ensure that data formats are suited to the modeling algorithm in use.

Duplicate records

As the data used in this study was sourced from the student data repository the issue of duplicate records did not arise.

Data entry errors

Data entry errors were corrected manually where possible by using the statistics tab in Rapid Miner® to review the example sets. are also at a minimum due to the design and setup of the original data warehouse from which these records were sourced. An example of a data entry error that occurred was that for a student the high school English grade which was set at 800 instead of 80. On reviewing the statistics tab in Rapid Miner® for data quality, characteristics and missing values it was noticed that the range was from 1-100 for this feature and the necessary steps were taken to correct this.

Outliers

Outliers were also checked using the statistics tab for each of the example sets to remove values that were out of bounds in the example set. There was no occurrence of any outliers.

Missing Values

Some of the pre-college attributes such as high school English, high school Math, CEPA and IELTS scores had missing values. The Impute Missing Values operator was used in Rapid Miner® to correct this. Impute Missing Values is a nested operator and k-NN was used in the subprocess to estimate missing values from the Example Set for each attribute. The value of k-NN was set at the default value of 5. The k-NN (k Nearest Neighbor) algorithm is very robust for missing values according to (Kotu and Deshpande, 2014). This operator was successful in replacing all the attributes that had missing values.

Transformation

The values of the campus and gender attributes were changed to numeric using a number for each campus code and 1 for Male and 2 for Female in the gender attribute. Normalization of the numeric data was also enabled using the Normalize operator in Rapid Miner®.

Feature selection

Methods used

Feature selection in data mining is one of the key factors that influences the success of predictive modeling. Filter and wrapper are the two types of feature selection methods. In the filter method attributes are selected based on a ranking of one of more criteria while the wrapper method is used during modeling to select attributes using forward selection or backward elimination.(Kotu and Deshpande, 2014) Dimension reduction methods is another feature selection option where attributes are merged to reduce the dimensionality of the dataset. Principal component analysis (PCA) is used for dimension reduction. According to (Shlens, 2014), the objective of using PCA is to identify the attributes that are most relevant and meaningful and hence filter noise. Hence the attributes with the maximum variability in the dataset are identified by transforming the original set of attributes into 'principal components that are not corelated to each other, collectively can define the variance in the dataset and can be related to the original attributes using the weightage factors. Feature selection can also be done by selecting the attributes that have a strong correlation to the predicted or target variable. Information gain or gain ratio measures the information exchanged between the attributes and the target variable. (Kotu and Deshpande, 2014)

The original dataset in this study was made up two different files, one that contained personal details and pre college information such as high school grades and common exam scores. The other file included information on courses taken registration status, absences and academic standing for each student. The combined dataset of these two files had 100 attributes before preprocessing was done.

Feature reduction was initially done based on honoring privacy of the students and hence all attributes that related to personal information and as such had no relevant to this study were removed. Using the context and knowledge of the environment of this study other attributes such as volunteering details and registration information was also removed from the dataset as they would have no bearing on the 'At Risk' prediction which is the focus of this study. The final set of attributes that were considered were 14 in number, details of this are provided below.

As the objective of this study is to predict the risk factors for students starting the program in Year1. It is important to ascertain the possibility of predicting 'At Risk' without considering the course grades that students had completed in the first semester of Year 1. Here the intention was to determine if students are 'At Risk' when they begin the first semester in the program. Hence the models were built on a dataset that did not include the course grades. The attribute 'Academic Standing' is the binominal label which identifies the examples as belonging to one of the two classes namely 'At Risk' and 'Good Standing'. Students who had a GPA of below 2.0 were considered as At Risk, this is in compliance with what is currently followed at the university on which this study is based.

Dataset attribute details

Dataset 1

Attribute Name	Attribute Description	Attribute Type	Comments
ID	Student ID	Polynominal	ID was **not** used included as a model attribute. ID was used a join attribute to combine two files for some models that were tested using course data

Gender	Change to the following values using VLOOKUP in MS Excel:
	M=1
	F =2
		Integer	

Campus	Campus codes were changed to numerical values using numbers. E.g.: Dubai campus was replaced with the value 1	Integer	
Cumulative GPA (CGPA)	Overall GPA	Real	CGPA was **not** included as a model attribute. Label was based on CGPA value
High school Average	High School overall grade	Real	
High School English	English grade	Real	
High School Math	Math grade	Real	
CEPA	Common exam score	Integer	
IELTS Band	IELTS Score	Real	
Absence %	Average attendance percentage	Real	
Academic Standing	**LABEL** At Risk (< CGPA of 2) or Good Standing (> CGPA of 2)	Binominal	

Table 3—2 Feature Set – Dataset 1

Dataset 2			
Attribute Name	Attribute Description	Attribute Type	**Comments**
Username	ID	Integer	**Not** included as a model attribute
C1_Quiz1	Course 1 Quiz 1 grade	Real	
C1_Quiz2	Course 1 Quiz 2 grade	Real	
C1_Project	Course 1 Practical grade	Real	
C1_Research Project	Course 1 Project grade	Real	
C1_FE	Course 1 Final Exam grade	Real	
C1_Final Grade	Course 1 Overall grade	Real	
C2_Quiz1	Course 2 Quiz 1 grade	Real	
C2_Quiz2	Course 2 Quiz 2 grade	Real	
C2_Pract 1	Course 2 Practical 1 grade	Real	
C2_Pract 2	Course 2 Practical2 grade	Real	
C2_Project	Course 2 Project grade	Real	
C2_FE	Course 2 Final Exam grade	Real	
C2_Final Grade	Course 2 Overall grade	Real	
C3_Quiz1	Course 3 Quiz 1 grade	Real	

C3_Practical	Course 3 Practical grade	Real	
C3_Project Output	Course 3 Project grade	Real	
C3_FE	Course 3 Final Exam grade	Real	
C3_Final Grade	Course 3 Overall grade	Real	
CGPA	Cumulative GPA	Real	
High School Average	High School Grade	Real	
CEPA	CEPA Entrance Exam score	Integer	
IELTS	IELTS English level	Real	
GPA_Rating	LABEL At Risk (< CGPA of 2) or Good Standing (> CGPA of 2)	Binominal	

Table 3—3 Feature Set – Dataset 2

Dataset 1 is used to identify the best classification models for predicting At Risk in the given dataset. Feature set selection has been done for both Dataset 1 and Dataset 2. The dataset 2 is used for determining accurate early prediction of At-Risk students.

Some models were tested using the course grade information from Semester 1 courses to ascertain relevance of the course performance data in predicting At

Risk. These models did achieve a good performance but since the focus of this research is predicting At Risk when students start their studies in Year 1 of the program and to determine how early risk can be predicted with the course performance details, only the datasets 1 and 2 were used for modeling in this study and course grade information was not included in the model attributes for Dataset 1

Data Imbalance – Sampling and Weighting

Sampling

He et al., (2009) defines data imbalance as a data set that has unequal distribution between the classes. Their study also suggests that the use of sampling techniques to balance the data improves classifier accuracy. Stratified sampling, Bootstrap sampling and Synthetic Minority Oversampling Technique (SMOTE) are the sampling techniques that were used in this study (He and Garcia, 2009). SMOTE uses synthetic samples to oversample the minority class rather than oversampling with replacement. (Chawla *et al.*, 2002) The Sampling stratified operator was used in Rapid Miner to sample the data ensuring the classes were equally represented. The number of examples in the sample was set to absolute. A macro was defined using the Extract Macro operator to count the values in the minority class and balance the classes. The balance data parameter was enabled, and sample size was set to the macro value. The bootstrap sampling has also been used with the data mining algorithms for sampling with replacement in this study. A macro has been enabled using the extract macro operator to count the minority class and set the sampling ratio.

Weighting

Using weighting to control the output of a classifier also help improve class imbalance.(Krawczyk, 2016). The Generate Weight Stratification operator in Rapid Miner® divides the weight specified in the 'total weight' parameter across all the examples. This operator ensures that the sum of the example weights for all labels, in this case 'At Risk' and 'Good Standing' are the same. The Extract Macro operator is used to create a macro that counts the number of examples. This was then enabled as the total weight parameter for generate weight stratification. This ensures that new examples coming in to the dataset will also be accounted for in the distribution of the weight for the example set. Sampling and weighting were enabled for each of the data mining models that were used in this study.

Modeling

As the focus of this study was student performance prediction a predictive data mining technique namely Classification was used. The selection of algorithms used was based on the findings of multiple studies in the literature review. Classification and predictive modeling using Naïve Bayes, Decision Tree, SVM, Neural Network, Logistic regression and Random Forest are used widely by researchers in data mining for prediction analysis.(Shahiri, Husain and Rashid, 2015; Aulck *et al.*, 2016). Over the past decade, these models or subsets of these models were predominantly used for predictive classification. Hence these models were chosen to determine the best classifier for predicting students at risk in the given dataset. Another feature that was adopted from current studies is the use of multiple classification algorithms to compare model performance measure values and determine the best classifiers for solving an educational issue.

Hence multiple classification algorithms were used with the models to compare performance and determine the best models for student ‘At Risk’ prediction. The graphic below summarizes the predictive classification techniques and corresponding algorithms that were used in Rapid Miner® to build the classification models for the data sets.

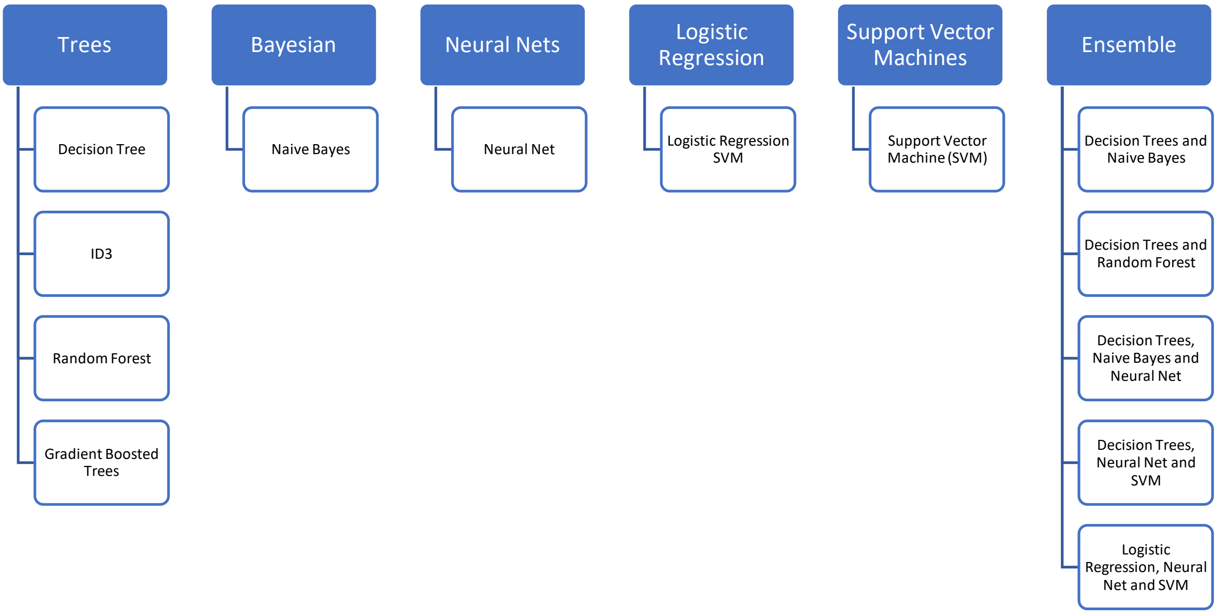

Trees
Decision Tree
ID3
Random Forest
Gradient Boosted Trees
Bayesian
Naive Bayes
Neural Nets
Neural Net
Logistic Regression
Logistic Regression SVM
Support Vector Machines
Support Vector Machine (SVM)
Ensemble
Decision Trees and Naive Bayes
Decision Trees and Random Forest
Decision Trees, Naive Bayes and Neural Net
Decision Trees, Neural Net and SVM
Logistic Regression, Neural Net and SVM

Figure 3-2 Classification Algorithms

Model Characteristics

Data mining tasks can be categorized as predictive or descriptive based on the objective. Predictive tasks focus on the prediction of a target or independent variable based on the other attributes of the data set known as explanatory or independent variables. In descriptive tasks the objective is to find patterns that summarize attribute relationships in the data. Clusters, trends and correlations are some of the examples for descriptive tasks.(An, Steinbach and Kumar, 2013).

“A model is the abstract representation of the data and its relationships in a given data set” (Kotu and Deshpande, 2014). According to Awad and Khanna, (2015) models provide a structure that summarizes the dataset for either prediction or description. Descriptive modeling generally applies unsupervised learning functions to produce patterns that explain the relationships and interconnections in the data. On the other hand, predictive modeling uses supervised learning functions to estimate future or unknown values of the target variable based on the related features of the independent variables. (Peña-Ayala, 2014). Typical steps in predictive modeling are as shown below:

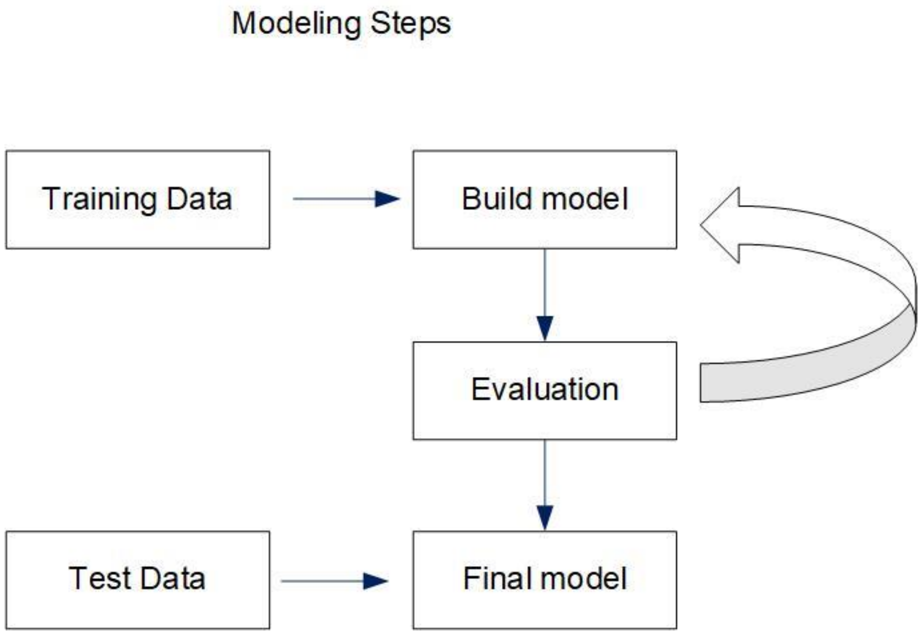


Figure 3-3 Modeling Steps (Kotu and Deshpande, 2014)

Classification and regression are the two types of predictive modeling techniques. Predictive modeling algorithms need access to training data to learn the model. A test data set is then used to predict the label based on the knowledge acquired from the training data. Hence classification is categorized as supervised learning.

Classification techniques or classifiers is the use of a systematic approach to build a classification model using a specific data set. Some examples of classifiers are Decision Tree, Naïve Bayes, Neural Networks, Rule Based, and Support Vector Machines (SVM). Each of these classifiers use a learning algorithm to build a model that is able to correctly learn and predict the relationship between the attributes set and the target variable or class. The effectiveness of the model is when it can correctly predict the label or class for unknown or new records in the data set. The model is trained on a subset of the data with known class labels. Evaluation of the model is the accuracy it can achieve on the test data (unknown class labels).

Model Evaluation

A classifier typically learns to build a model on the training data. This knowledge is then used on the test data to classify or label unknown instances. The objective of evaluating a classifier is to measure its performance quality on the test data. (Aggarwal, 2015). It is important therefore that the samples used for the training and testing data sets are valid and representative of the data set used. k-fold cross validation is used to partition the data into k independent subsets. Of this k-1 subsets are used to build the model and the kth subset is used as the test data. This process is continued iteratively until k different models are achieved and the results of these k models are then combined used either average or voting. The advantage of using k-fold cross validation is that each record is used only once in the test set. (Larose, 2015). A 15-fold cross validation was utilized to balance the training and testing subsets of the data for all the models in this study using stratified as the sampling setting in the cross-validation operator. The 15-fold cross validation process divides the data into 15 roughly equal parts. Classification is then done on each part (test data) by using the remaining 14 parts to train the model. The evaluation of each part finally results in 15 different values for each part that is tested for the performance measures. This is then averaged for the final result. (Márquez-Vera *et al.*, 2016). Using cross validation nullifies the danger of overfitting which happens when the model memorizes the learning from the training data. An overfitted model will thus underperform on the training data. The classification process consists of three main phases namely training, validation and testing. An unbiased evaluation of the trained model happens in the validation phase. (Tharwat, 2018). Accuracy, f-measure, kappa coefficient, AUC, precision and specificity and sensitivity or recall are some of the ways in which performance has been evaluated in other studies. As the dataset was not balanced between the positive and negative class, performance evaluation measures were selected based on the recommendations of other studies on imbalanced data. He et al, (2009) states that in the case of imbalanced learning a singular assessment criterion such as the conventional accuracy or error rate will not be sufficient in terms of classifier performance evaluation. Class distribution will have to be taken into consideration when deciding on performance measures if the dataset is imbalanced. Class distribution is the ratio between the positive and negative samples which is represented by the left and right columns of the confusion matrix. A classification measure such as accuracy or its complement classification error rate that uses both columns of data in the confusion matrix will hence be sensitive to imbalanced data and cannot be used on its own to measure performance. Geometric Mean on the other hand can be used to measure performance of imbalanced data although its value is computed from both columns of data because changes in the class distribution cancel each other. The other drawback with accuracy is that accuracy can be the same for two classifiers but the values for the confusion matrix in terms of correct and incorrect positive and negative values will not tally. Sensitivity and specificity use values from the same column in the confusion matric and hence are insensitive to imbalanced data and can be used as a performance measure for the same. (Tharwat, 2018). Other measures such as AUC values from the ROC curves and accuracy was used to measure the performance in the machine learning models namely logistic regression, random forest and K-nearest neighbors on a balanced dataset using random sampling. Aulck et al, (2016) and Zou et al., (2016), argue on the practice of using AUC values, which is generally considered as a reliable performance metric for binary classification. The argument is that it may not always be the case that a high value of AUC correctly reflects a good classification performance due to the presence of trash negative values. These values are difficult to distinguish and although it increases the AUC value, precision and recall values can suffer. F-score or f-measure is used along with AUC in this study to counter this issue, high values of ROC generally result in high values of f-measure (Aulck *et al.*, 2016; Zou *et al.*, 2016). Costa *et al.* (2017) also used f-measure to evaluate the effectiveness of educational data mining models to predict student failures. Cohen’s Kappa is also used to measure the accuracy of predictions in classification models.(Ben-David, 2008). Kappa scores represent how close the accuracy of the model is to the actual values by denoting how much better the model is to a chance prediction(Baker *et al.*, 2015). Paul Akangah *et al*. (2018) defines models that achieve a kappa value of over 0.5 as being in substantial agreement with the data.

The next section provides a brief description on how these performance measures are used for model evaluation in this study.

The effectiveness of a classification model can be assessed using tools such as the confusion matrix, ROC curves or lift charts. Confusion matrix is also known as a truth table and is usually arranged as 2 x 2 matrix in which the predicted classes are placed horizontally in rows while the actual classes are placed vertically in columns.(Kotu and Deshpande, 2014). A binomial classification has been used in this study with the following classes ‘At Risk’ and ‘Good Standing’. Classification models can be evaluated using the information from the confusion matrix table shown below.

		Actual Class (Observation)	
		Y	N
Predicted Class (Expectation)	Y	**TP** True positive (Correct Result)	**FP** False Positive (Unexpected

			Result)
	N	**FN** False negative (Missing result)	**TN** True Negative (Correct Absence of result)

Figure 3-4 Confusion Matrix (Kotu and Deshpande, 2014)

TP or True positive is when the predicted class is a ‘Y’ and the actual class is also a ‘Y. **TN or True Negative** on the other hand is when the predicted class is a ‘N’ and the actual class is also a ‘N’.

FP of False positive is when the predicted class a ‘Y’ and the actual class is a ‘N’, **FN or False Negative** is when the predicted class is a ‘N’ and the actual class is a ‘Y’ A good classification algorithm will have minimum values for FP and FN with for FP and FN being equal to 0 in a perfect classifier.

Using the above information, we can calculate the following parameters to evaluate the effectiveness of the predicted model.

Sensitivity is the ability of a classifier to predict a ‘Y” for all for examples that are a ‘Y’. without missing any examples that has a ‘Y’ value thus eliminating false negatives. In reality, a confusion matrix will always have some value representing false negatives. Sensitivity is calculated as the ratio or percentage of $\frac{TP}{TP+FN}$

Specificity measures the ability of a classifier to reject all ‘N’ values. A prefect classifier will have a FP value of zero. Specificity is calculated as the ratio or percentage of $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$

Relevance of the examples that are labelled correctly leads us to identify the measures of precision and recall. **Precision** measures the examples that are actually positive in the group that is defined as the positive class. It is calculated as follows $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$. **Recall** is the percentage of correct labels correctly predicted by the classifier. It is calculated as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ which is identical to calculating sensitivity. Higher precision results in lower false positive errors and classifiers that have a larger recall value have fewer examples of the positive class misclassified as negative. In binary classification the rare class is defined as the positive while the major class in defined as negative. **F1 measure** is the harmonic mean between recall and precision and hence tends to be closer to the smaller of the two values. A high value of the F1 measure ensures that both recall and precision are relatively high. (An, Steinbach and Kumar, 2013)

Accuracy measures how well the classifier is able to select all examples with a ‘Y and reject those with a ‘N”. A classifier with a 100% accuracy would have both FP and FN values at zero. The following ratio or percentage is used to calculate accuracy $\frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}$. **Error** is calculated as (1- accuracy). Accuracy treats all classes as equally important and hence is not a good measure for imbalanced data. Data is imbalanced or skewed when one class has more representation and outnumber the other(s) in terms of the number of examples in that class. Machine learning algorithms operate on the assumption that the classes are roughly balanced. When this is not the case in real life situations the bias is towards the majority class which could result in more prediction errors in the minority class. The minority class although rare and insignificant in terms of the number of examples could provide useful and important information which could be impacted negatively by this bias. **Geometric mean (GM)** is a better measure to use in this case as it is insensitive to imbalanced data. GM aggregates both sensitivity and specificity measures. GM = $\sqrt{sensitivity \times specificity}$

A **receiver operating characteristic (ROC)** curve depicts the relationship between the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. (Krawczyk, 2016)

TPR = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (% of correctly classified positive samples)

FPR = $\frac{\mathrm{FP}}{\mathrm{FP+TN}}$ (% of incorrectly classified negative samples)

The **area under the curve (AUC)** is basically the area of the right-angled triangle with a breadth and height of 1 which is 0.5. AUC for a perfect classifier is 1.(Kotu and Deshpande, 2014). A sample of a ROC curve for a classification model used in this study is shown below:

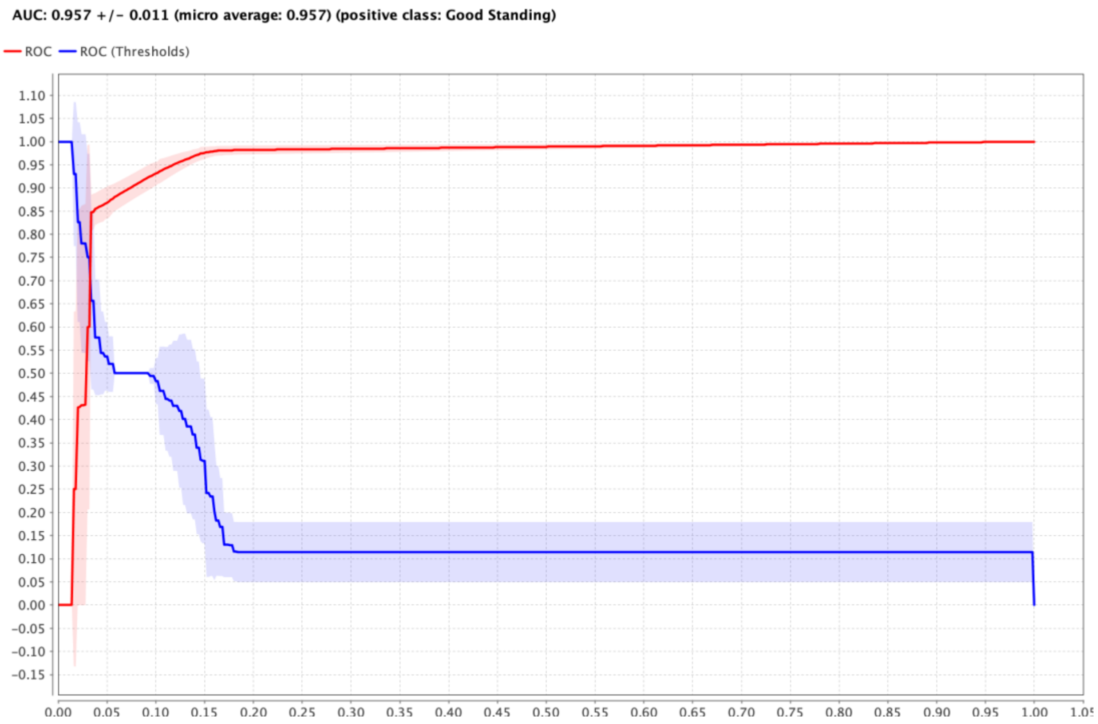
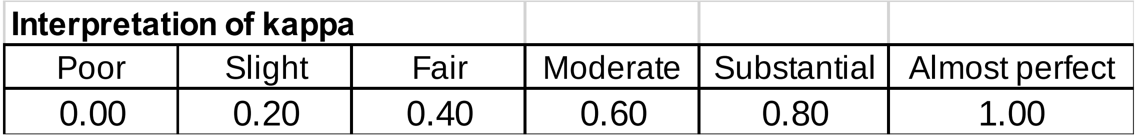


Figure 3-5 ROC Curve Sample

A good classification model should be located close to the upper left corner of the diagram. (An, Steinbach and Kumar, 2013). Although ROC is considered as the ‘gold standard’ for imbalanced data model evaluation, the ROC and AUC values only indicate the ranking power of positive prediction probability. So very high values of AUC can still have low values of precision and recall due to negative samples. It is recommended that f-measure is used along with AUC for better model evaluation in the case of imbalanced data.(Zou *et al.*, 2016)

Cohen’s Kappa can also be used to measure accuracy of a classifier. While measure the degree of agreement it subtracts the portion of the counts that could be ascribed to chance. The values for this measure ranges from -1 which suggests total disagreement to 1 which is total agreement. A value of 0 for the Cohens Kappa statistic indicates a random classification. (Ben-David, 2008)



Interpretation of kappa					
Poor	Slight	Fair	Moderate	Substantial	Almost perfect
0.00	0.20	0.40	0.60	0.80	1.00

Figure 3-6 Kappa Scores Interpretation (Viera and Garrett, 2005)

Evaluation Methodology

The data used in this study is imbalanced with more occurrences of ‘At Risk’ which is the majority class and comparatively fewer instances of the minority class ‘Good Standing’. Although the focus of this study is to determine the students being labelled as At Risk, the majority class, it is established from the literature reviewed on class imbalances that effectiveness of the model should be based on measurements for both classes. Hence accuracy values from the confusion matrix alone cannot be used to determine the efficacy of the model. The models built in this case study will be assessed using the following measures:

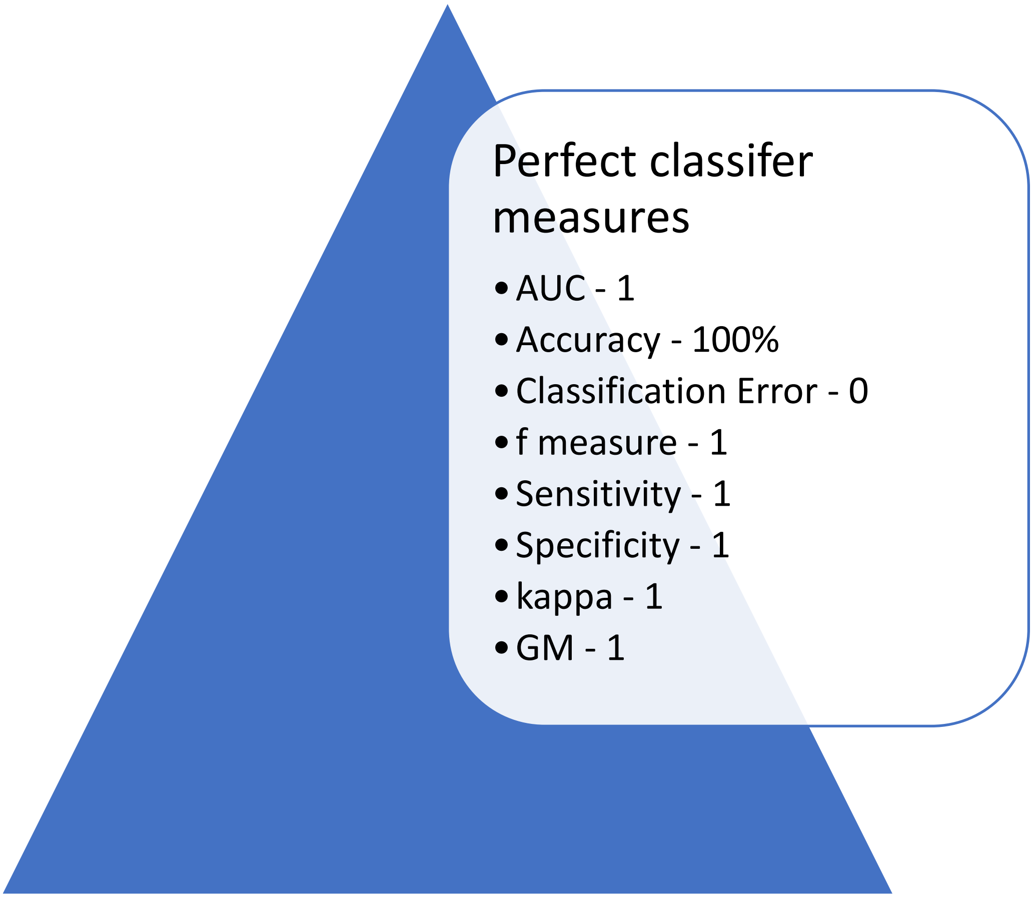


Figure 3-7 Classifier Performance Measures

Precision values will also be discussed for each model

Deployment

Models were built using Rapid Miner Studio® which is a visual data science program used for designing prototypes and can be used for model validation. The previous experience with this software, its extensive capabilities, access to Rapid Miner® Studio as it offers free academic licenses for students and the support provided with tutorials, rapid miner community access and documentation contributed to its selection as the data science and machine learning modelling software (RapidMiner Studio 9.2', 2019).

A brief description of the classification models used in this case study along with the parameter changes that were introduced in its practical implementation in Rapid Miner® is detailed below:

Decision Trees

Decision tree is widely used as a classifier and it works well with both numerical and categorical data. Classification of data starts at the root of the tree also called the decision node and continues onward resulting in internal and leaf nodes. During induction of a Decision Tree, each internal node corresponds to a splitting decision that partitions the domain of one or more attributes of the data set. This continues until a leaf or end node is reached which identifies the class label. (Aggarwal, 2015, chap. 2). The splitting decision is based on the homogeneity of the data. The measure of impurity is maximum when all the possible classes are represented equally and zero when the data set only represents one class. Entropy or Gini index are measures that meet this criterion and thus used for Decision Tree building. Information for an attribute is computed as the weighted sum of its component entropies. Information gain is the information before the split minus the information after the split. The disadvantage of this method is that it is biased towards choosing attributes with a large number of values as root nodes. This can be avoided by using gain ratio which overcomes this by taking into account the number of branches that would result before making the split, so the intrinsic information of a split is accounted for. Features that have a high uncertainty will not be automatically selected as using it for splitting will offer low gains . (Kotu and Deshpande, 2014)

Advantage of using Decision Tree as a classifier include the simplicity of the model, computationally inexpensive thus providing a low cost high speed option for classification.(Karimi-Alavijeh, Jalili and Sadeghi, 2016). A good classification models should have low training and generalization errors. Model underfitting occurs when the training and generalization error rates are large with a small sized tree. This occurs because the model has still the learn the complete tree structure. A model that fits the training data too well will have a high generalization error and this will lead to overfitting. Reducing model complexity is one way to deal with overfitting. Pre-pruning to halt the growth of the Decision Tree before it reaches its full size is one approach to counter overfitting. Disadvantage of this method is the danger of pruning too early which could lead to underfitting. Post-pruning requires the tree to reach its maximum size before pruning occurs bottom-up. Here trimming occurs by replacing a sub tree with a leaf node with the class label of the majority of records associated with the subtree. Pruning in this case is terminated when no further improvement is observed. Post-pruning usually leads to better results when compared to pre-pruning.(An, Steinbach and Kumar, 2013)

Decision Tree modeling parameters that were used in Rapid Miner®

Parameter	**Default Value**	**Changed value**
Criterion	gain ratio	
Confidence	0.1	0.3
minimal gain	0.01	0.1

Pre and Post pruning	Both options are enabled	

Table 3—4 Decision Tree parameters

Criterion decides the criteria used for selecting the attributes for splitting. Information gain calculates the entropy all the attributes and the one with the least entropy is selected for the split. Gain ratio which is the default value for criterion is a variation of information gain that also takes into account the number of splits for the attribute. An attribute with a large number of splits will have low gain and hence will not be selected for splitting the data set.

Confidence parameter setting is used to calculate the pessimistic error calculation of pruning. When pruning is applied after the tree is built some branches will be replaced by leaves based on this value.

Minimal gain – When splitting a node, the gain is calculated and compared to the minimal gain value. Splitting of the node occurs only if its gain value is higher than the minimal gain parameter value. Higher values of minimal gain will thus result in smaller trees.

The values of confidence and minimal gain was varied, and the best model performance was obtained at confidence =0.3 and minimal gain = 0.1. All other values for Decision Tree remained unchanged.

ID3 (Iterative Dichotomiser 3)

ID3 is an algorithm for decision tress that was invented by Ross Quinlan. It is the precursor to the C4.5 algorithm. The basic function of this algorithm is to build a Decision Tree from a fixed set of examples and using this tree information to classify future samples. The criterion parameter is used to select the feature selection heuristic that helps ID3 determine which attribute goes into the decision node. The information gain (used as the selection criteria) is calculated for each attribute and the one that holds the highest information gain is selected as the root of the tree. Based on the attribute values branches are generated and this process continues recursively until the sub dataset in each branch has the same class label. This tree is then used to classify the new examples. It should be noted that ID3 is prone to overfitting with smaller datasets.

((Yang, Guo and Jin, 2018)

The parameters for ID3 classification modelling in Rapid Miner was retained as default values. Altering the minimal gain values did not improve the model performance results.

Parameter	**Default Value**
Criterion	gain ratio
minimal gain	0.01

Table 3—5 ID3 parameters

Decision Tree and its variations namely ID3 and CHAID models were compared on the datasets. ID3 had the better performance measure values in comparison to Decision Tree and CHAID for the models using course grades. This model was not investigated further as it does not work with numeric values.

Naïve Bayes

Naïve Bayes algorithm works on the probabilistic relationship between the class label (predictor) and the factors(attributes) to build a classification model. The ‘naïve’ tag comes from the characteristic of the algorithm to naively assume that attributes are independent from one another. This may not always hold true.

Naïve Bayes calculates the probability from the data set and hence if sampling is used it is imperative that data being mined is truly representative of the population. Model building is quite simple and includes creating a lookup table of probabilities. Incomplete or missing values in the training set and the assumption of independent variable are the two main drawbacks of this algorithm, Laplace correction is the parameter used to counter missing attribute values in the training set. This sets small default probabilities for missing values instead of the zero that causes misleading results.(Kotu and Deshpande, 2014) The Naïve Bayes algorithm was used in this study with the default setting for the Laplace parameter

Parameter	**Default Value**
Laplace correction	enabled

Table 3—6 Naive Bayes Parameters

Logistic regression

Logistic regression uses the logit function to predict the probability of an event occurrence based on categorical attributes or predictors. It is a modeling technique that discovers the relationship between the input variables and a categorical target variable.(Raju and Schumacker, 2015). Logistic regression allows you to fit the data to a nonlinear curve when using a discrete target variable.(Deshpande, Bala ; Kotu, 2018)

Logistic regression modelling in Rapid Miner uses the Logistic Regression learner operator. This learner is based on internal Java implementation of the myKLR by Stefan Rüping.(Rüping, 2003). *myKLR* is based on the code of *mySVM* and hence the format of example files, parameter files and kernel definition are identical.(Rüping, 2000)

This operator supports various kernel types, the default setting is dot. The kernel radial with a kernel gamma setting of 1.0 (default setting) was used in the study to build a Logistic Regression model. The complexity constant parameter, 'C' that sets the boundary for misclassification errors was set to 100. Parameters were changed as altering the kernel and C values resulted in better model performance as indicated by the evaluation measures.

Parameter	**Default Value**	**Changed setting**
Kernel	dot	radial
kernel gamma	1.0	
C	1.0	100

Table 3—7 Logistic Regression Parameters

Neural Nets

Neural networks are composed of nodes that are interconnected and directed link. The functional relationship between the input variables and the target variable is built similar to the biological process of a neuron. The network has the input nodes or units and the output nodes which is the last layer. Inputs from previous layers are connected to the nodes from hidden layers resulting in a complex combination of input values. (Deshpande, Bala ; Kotu, 2018).

The Rapid Miner® implementation of this classifier learns the model using a feed forward neural net. A back-propagation algorithm is used to train the neural net. A feed forward neural network has no cycles or loops and information moves only in one direction from the input to the output node via hidden nodes if any. The back-propagation algorithm consists of two phases propagation and weight update of the connectors. This process reduces the error functions and results in a target state.

The structure of the neural network model implementation was altered by introducing three hidden layers of size 14, 10 and 6. The learning rate and momentum parameters were changed from default to 0.4 to achieve optimum performance levels.

Parameter	**Default Value**	**Changed setting**
hidden layers	0	3 layers of size 14, 10 and 6
learning rate	0.01	0.4
Momentum	0.9	0.4

Table 3—8 Neural Net Parameters

Support Vector Machines (SVM)

SVM is a discriminant classification function that takes a data point and assigns it to one of the different classes of the classification task. SVM is one of the most popular learning approaches for supervised learning. Some of the main features that contribute to its popularity are as follows:

Although it requires all the training data to be stored in memory during the training phase when it is learning the model parameters thereafter it depends on only a subset of the training data called support vectors to make future predictions. SVM maps the data using pre-defined kernel functions to learn the linear separator between the classes. Kernel settings play a key role in achieving optimal model performance. SVM also places an additional constraint to ensure optimization which is that the hyperplane should be of equal and maximum distance between classes. This ensures better prediction ability on new examples.(Awad and Khanna, 2015 Chapter 3)

There are two type of classification in SVM. Hard margin classification is inflexible in allowing misclassifications and works better with linear data that allows for rigid lines of separation between the classes. This method is also prone to outliers. Soft classification on the other hand is more flexible and is a good balance between the width between the boundaries and limiting the number of misclassifications. Soft margin classification can be controlled using the ‘C’ parameter in SVM. (Awad and Khanna, 2015; Géron, 2017)

The SVM parameters were varied for kernel and C and it was found that the radial setting for kernel with C=100 provided the best model performance

Parameter	**Default Value**	**Changed setting**
Kernel	dot	radial
kernel gamma	1.0	

C	0	100

Table 3—9 SVM Parameters

Ensemble Models

Ensemble modeling uses more than one classifier to predict an outcome. Using ensemble models reduces the generalization error and so long as the models used are diverse and independent prediction errors should reduce. Even though the model contains multiple base models it still functions as one model.(Deshpande, Bala ; Kotu, 2018). Ensemble methods use a set of base classifiers on training data. Classification is achieved by using a voting system on the predictions made by each classifier(An, Steinbach and Kumar, 2013)

Bagging and Boosting are two types of ensemble methods. Bagging also known as bootstrap aggregating repeatedly samples (with replacement) where all records have equal probability of selection and each sample is the same size as the training set. These samples are called bootstrap samples. Classification is then done for each sample and a prediction is recorded. The class with the most votes is selected as the bagging ensemble prediction for classification. Bagging works better with unstable models such as Decision Tree or neural networks, by reducing the variance but can actually worsen model performance for stable classifiers. In Boosting the same classification model is applied to the samples but records that are misclassified are given a higher weight in each iteration. The final boosted classifier is the weighted average of the base classifiers. Boosting helps to reduce bias and variance in the predictive models. However if a stable classifier is used then boosting can increase the variance. (Kotu and Deshpande, 2014)

Bagging and Boosting with Adaboost operator were applied with each of the models in Rapid Miner®. Default settings were retained for both operators with the Adaboost iteration operator set to 10.

Random Forests

Random Forests is an ensemble that uses bagging and is specifically designed for Decision Tree classifiers. Multiple Decision trees are generated based on independent random vectors. The prediction is based on the combined prediction of the various trees. Sampling is done with replacement by choosing a random number of samples from the original training set.(An, Steinbach and Kumar, 2013)

Random Forests was used a classification model in this with default values for the parameter 'number of trees. Pruning was enabled with confidence and minimal gain parameters set to default.

Parameter	**Default Value**
number of trees	100
Criterion	gain ratio
pre pruning and post pruning	enabled
confidence	0.1
minimal gain	0.1

Table 3—10 Random Forest Parameters

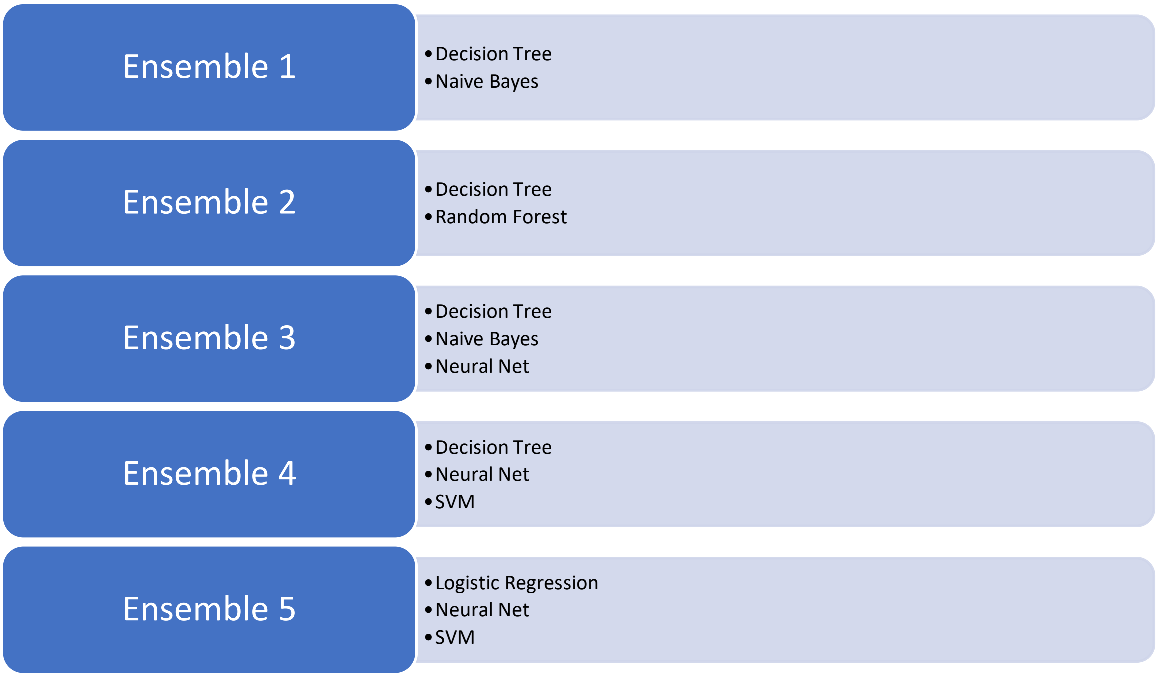
Gradient boosted trees

Gradient boosted trees are another ensemble model and uses boosting for classification. A single tree model is improved by using weighting to train the classification model. With each iteration examples are reweighted based on their previous prediction. The final model is a weighted sum of all the models that are created.

Gradient boosted trees is used as classification model in this study with the default parameters.

Ensemble model with multiple classifiers

The nested operator Vote was used to build ensembles with multiple classifiers. The following ensembles were used in this study to evaluate classification performance on the dataset:



Ensemble 1
- Decision Tree
- Naive Bayes

Ensemble 2
- Decision Tree
- Random Forest

Ensemble 3
- Decision Tree
- Naive Bayes
- Neural Net

Ensemble 4
- Decision Tree
- Neural Net
- SVM

Ensemble 5
- Logistic Regression
- Neural Net
- SVM

Figure 3-8 Ensemble Models used in this study

All base classifiers were used with the optimal parameters identified when they were used as individual classifiers in this study.

Chapter 4 Results

The following section will elaborate on the results of this study and provide the answers to the research questions. Research Questions

The primary goal of this research study is to evaluate a data mining approach for performance prediction and in doing so identify the best models in terms of performance measure values. The research question that this section will cover is

‘What are the most effective and efficient data mining algorithms for predicting students at risk using the given dataset?’

Classification is the data mining approach that is widely used for performance prediction in educational data mining. (Adejo and Connolly, 2018; Burgos *et al.*, 2018; Kumari, Jain and Pamula, 2018; Mhetre and Nagar, 2018). Multiple classification algorithms were used with the models on the two datasets that are used in this study based on the methodology used by the majority of researchers whose work is based on student performance prediction. Decision tree, Logistic Regression, Naïve Bayes and SVM have been commonly used for predicting student performance (Refer Appendix 1). Neural Net and ID3 has also been used by a few researchers with moderate success.(Agnihotri and Ott, 2014; Conijn and Zaanen, 2017; Kumari, Jain and Pamula, 2018).

In the following section an analysis of the results for the two datasets based on the techniques used such as sampling, weighting and ensemble models will be provided for the performance evaluation measures such AUC, geometric mean, f-measure, accuracy and Cohen’s kappa. Research in this area has shown that if a learning model achieves a good result with a particular measure this does not have to repeat if another measure is used with the same dataset. (Ferri, Hernández-Orallo and Modroiu, 2008) Keeping this in perspective in addition to the fact that there was a slight imbalance in the dataset multiple measures were used in this study to evaluate classifier performance. Accuracy is used a measure in a lot of the work that was reviewed but the importance of using other measures when data is imbalanced was also emphasized in more than one study. Hence the more robust measures such as geometric mean, AUC, f-measure and kappa scores were used along with accuracy for each of the models. Based on literature reviewed the model performance is generally considered as high if the values are 0.85 or above and closer to 1. A very high AUC value does not always mean that the model is performing well so this measure alone should not be used to determine the performance of a model. In this research AUC will be used along with the accuracy, geometric mean (G-mean), f-measure and kappa scores to consider if the model has high values in all these performance measures. Sensitivity and Specificity also helps to establish the true positive and true negative values which will be represent by geometric mean (G-mean). Higher precision and recall values also indicate low errors in false positive and false negative model values.

A comparison of the model performance using the various algorithms is provided in the table below. Models that have achieved an AUC score of above 0.8 and high scores in the other performance measures will be selected as the best performing models. Model performance details are provided below:

Model Comparisons									
	LR	**Bagging SVM**	**Boosting SVM**	**SMOTE SVM**	**NN SS**	**NN**	**NB BS**	**NB SS**	**DT**

AUC	0.953	0.944	0.942	0.942	0.817	0.798	0.747	0.741	0.662
Accuracy	0.908	0.911	0.911	0.910	0.764	0.742	0.696	0.702	0.694
G-mean	0.905	0.909	0.909	0.909	0.750	0.734	0.687	0.691	0.59
kappa	0.814	0.82	0.819	0.818	0.515	0.478	0.381	0.391	0.589
f-measure	0.895	0.899	0.898	0.899	0.715	0.697	0.648	0.650	0.512
Specificity	92.99	92.95	92.71	92.24	83.81	80.61	74.36	76.25	95.77
Sensitivity	88.12	88.83	89.06	89.54	67.07	66.77	63.54	62.53	36.29
Precision	90.96	91	90.78	90.31	77.04	74.12	66.45	67.73	87.33
False Positive FP	9.93	10	10.33	11	22.93	41.2	54.5	50.5	9.00
False Negative FN	13.4	12.6	12.33	11.8	37.13	56.2	61.7	63.4	107.80
Classify error	0.092	0.089	0.089	0.089	0.236	0.258	0.304	0.298	0.31

Table 4—1 Model Performance Comparison

Logistic regression model is the only original model without sampling, weighting or ensemble operators such as bagging or boosting that achieved high values in all the performance measure. The high sensitivity/recall and precision values also indicate that the predictions are reliable. SVM with the boosting, bagging and SMOTE operators also performed with high values for the different measures. The significance of these models is that the performance has high values for all the measures.

Model performance details for the algorithms used in this study is provided below with a short summary of the performance results for each of the models.

Decision Tree performance results

Decision tree (dataset excluding course grades)

	Decision Tree	Bootstrap Sampling	Bagging	Boosting	Weighting	Stratified Sampling
AUC	0.662	0.661	0.661	0.660	0.660	0.658
Accuracy	0.694	0.694	0.693	0.694	0.693	0.691
G-Mean	0.590	0.591	0.626	0.590	0.589	0.589
Kappa	0.589	0.591	0.591	0.590	0.589	0.589
f-measure	0.512	0.513	0.513	0.512	0.511	0.510
Specificity	95.77	95.67	95.20	95.77	95.49	95.01
Sensitivity	36.29	36.46	36.64	36.29	36.29	36.46
Precision	87.33	87.11	85.97	87.33	86.71	85.61
False Positive FP	9.00	9.20	6.80	9.00	9.60	10.60
False Negative FN	107.80	107.50	71.47	107.80	107.80	107.50
Classify error	0.306	0.306	0.307	0.306	0.307	0.309

Table 4—2 Decision Tree results

The performance measures for all the models are almost the same for all the Decision Tree algorithm. The model that uses bagging has a slightly higher geometric mean value due to a higher value of specificity as compared to the other models.

All the models have high specificity values and low recall values which results in higher values for false negatives. The Decision Tree model using bagging has the lowest value in errors for both false positives and false negatives.

ID3

ID3 could not be used as the dataset contained numerical values. When using ID3 with the course grades and nominal values the results showed high values for the performance measures. Since using a nominal to numeric operator in Rapid Miner®, is not recommended for accuracy ID3 was excluded from the study.

Naïve Bayes Performance Measure Values

Naïve Bayes (dataset excluding course grades)

	Bootstrap Sampling	Stratified Sampling	Naïve Bayes	Weighting	Boosting	Bagging
AUC	0.747	0.741	0.741	0.741	0.736	0.701
Accuracy	0.696	0.702	0.700	0.690	0.700	0.598
G-mean	0.687	0.691	0.69	0.704	0.69	0.598
Kappa	0.381	0.391	0.388	0.372	0.388	0.195
f-measure	0.648	0.650	0.649	0.651	0.649	0.520
Specificity	74.36	76.25	75.82	75.97	75.82	60.48
Sensitivity	63.54	62.53	62.71	65.26	62.71	59.04
Precision	66.45	67.73	67.41	64.99	67.41	59.25
False Positive FP	54.5	50.5	51.4	39.73	51.4	168
False Negative FN	61.7	63.4	63.1	39.2	63.1	138.6
Classify error	0.304	0.298	0.300	0.310	0.301	0.402

Table 4—3 Naïve Bayes results

The performance measure values for all the Naïve Bayes models are comparable in terms of their performance measure values. The model using bootstrap sampling achieving a marginally higher value of AUC. The kappa scores indicate only a slight agreement and overall the performance measure values are in the lower range. The values for specificity, sensitivity or recall and precision are in the lower range hence suggesting that values of false positives and false negatives are high. The model using weighting have the lowest errors of false positives and false negatives. The values for all the models are comparable and there is no one model that outperforms the others in terms of model performance or minimal errors.

Neural Net Performance Measure Values

Neural Net (dataset excluding course grades)

	Stratified Sampling	Neural Net	Weighting	Bagging	Boosting	Bootstrap Sampling	SMOTE
AUC	0.817	0.798	0.796	0.789	0.789	0.771	0.758
Accuracy	0.764	0.742	0.737	0.727	0.723	0.723	0.712
G-mean	0.750	0.734	0.733	0.703	0.700	0.721	0.697
Kappa	0.515	0.478	0.467	0.435	0.427	0.444	0.410
f-measure	0.715	0.697	0.703	0.658	0.654	0.689	0.653
Specificity	83.81	80.61	76.28	83.3	82.64	75.54	78.69
Sensitivity	67.07	66.77	70.44	59.34	59.27	68.86	61.81
Precision	77.04	74.12	70.83	74.11	73.29	69.5	70.81
False Positive FP	22.93	41.2	50.4	23.67	36.9	52	45.3
False Negative FN	37.13	56.2	50	45.87	68.9	52.7	64.6
Classify error	0.236	0.258	0.263	0.273	0.277	0.274	0.288

Table 4—4 Neural Net results

The Neural Net models have better values for the performance measures when compared to Decision Tree and Naïve Bayes models. The model with stratified sampling has the highest values for all the performance measures. The kappa scores have also improved from the Naïve Bayes models depicting a fair agreement. The Neural Net models achieved a good score for specificity in some of the models while the precision and recall/sensitivity values were average. False positive and False Negative errors were lowest in the model that uses stratified sampling. This model also achieved the highest specificity and precision scores with comparably the lowest value for classification errors

SVM Performance Measure Values

SVM (dataset excluding course grades)

	Bagging	Boosting	SMOTE	SVM	Stratified Sampling	Weighting	Bootstrap Sampling
AUC	0.944	0.942	0.942	0.913	0.883	0.835	0.819
Accuracy	0.911	0.911	0.910	0.810	0.834	0.712	0.784
G-mean	0.909	0.909	0.909	0.819	0.835	0.717	0.781
Kappa	0.82	0.819	0.818	0.635	0.671	0.431	0.563
f-measure	0.899	0.898	0.899	0.818	0.816	0.684	0.767
Specificity	92.95	92.71	92.24	72.43	85.89	65.31	80.62
Sensitivity	88.83	89.06	89.54	92.67	81.09	78.63	75.59
Precision	91	90.78	90.31	73.49	82.28	70.93	76.15
False Positive FP	10	10.33	11	117.2	20	73.7	41.2
False Negative FN	12.6	12.33	11.8	24.8	21.33	36.2	41.3
Classify error	0.089	0.089	0.089	0.19	0.166	0.288	0.216

Table 4—5 SVM results

SVM with the ensemble operators namely bagging and boosting had high scores for the AUC performance measure. Although high AUC values do not always indicate high model performance it can be argued that since all the other performance measures such as geometric mean (G-mean), f-measure, accuracy and kappa scores also registered high values that SVM with Bagging or Boosting and SVM with SMOTE are the more efficient models for this dataset. As the kappa scores are above 0.8 for the majority of the models, this indicates a substantial agreement.

The SVM models using the ensemble operators such as Bagging and Boosting, and the SMOTE sampling technique achieved high values for specificity, sensitivity/recall and precision. The values were low in these models for the false positive, false negative and classification errors suggesting that the models are highly capable of making the correct predictions

Logistic Regression Performance Measure Values

Logistic Regression (dataset excluding course grades)

	Logistic Regression	**SMOTE**	**Bagging**	**Bootstrap Sampling**	**Stratified Sampling**
AUC	0.953	0.947	0.935	0.863	0.821
Accuracy	0.908	0.891	0.881	0.790	0.752
G-mean	0.905	0.889	0.877	0.778	0.727
Kappa	0.814	0.779	0.759	0.569	0.486
f-measure	0.895	0.877	0.863	0.748	0.685
Specificity	92.99	90.45	90.92	85.65	86.21

Sensitivity	88.12	87.41	84.63	70.58	61.23
Precision	90.96	88.03	88.16	79.68	78.09
False Positive FP	9.93	13.53	12.87	30.5	29.1
False Negative FN	13.4	14.2	17.33	182.1	65.6
Classify error	0.092	0.109	0.119	0.210	0.248

Table 4—6 Logistic Regression results

Logistic regression could not be used with weighting or Adaboost (boosting operator)

as it does not recognize the weights. Performance levels of the Logistic Regression models were improved by changing the default values of kernel and C to radial and 100 respectively.

The original Logistic Regression Model along with the models using SMOTE and Bagging had high values for AUC. As these models also have registered high scores for other performance measures it nullifies the contra indications of an inflated AUC. The kappa scores also indicate moderate to substantial agreement for these models. The Logistic Regression models using sampling both bootstrap and stratified have much lower performance measure values with fair to moderate agreement kappa score values. The Logistic Regression models namely the original model and the models using SMOTE and Bagging recorded high values for specificity, sensitivity/recall and precision. In the case of the error readings such as classification errors, false positive and false negative low values were recorded for these models indicating a fairly accurate model in performance prediction

Ensemble Models

The use of Ensemble models for classification has been reported in comparatively fewer studies but the common factor in these studies is the high-performance values that ensemble models achieve. Hence to evaluate the best performing classification model for the datasets in this study it was important to consider ensemble models using bagging, boosting, Gradient Boosted trees, random forest and the vote operator with multiple classifiers. The research question that this section will answer is ‘How do ensemble models perform in comparison to the models using individual classifiers’

The comparative results of the ensemble models in this study is given below.

Ensemble Models						
	GB	**LR, SVM and NN**	**DT, NN and SVM**	**LR, NN and SVM-SS**	**DT and NB**	**RF**
AUC	0.937	0.917	0.885	0.826	0.717	0.689
Accuracy	0.870	0.913	0.793	0.814	0.690	0.692
G-mean	0.871	0.909	0.760	0.807	0.584	0.592
Kappa	0.739	0.822	0.566	0.62	0.332	0.338
f-measure	0.858	0.900	0.721	0.784	0.506	0.516
Specificity	86.22	93.41	92.89	85.27	95.34	94.87
Sensitivity	88.07	88.41	62.1	76.42	35.87	37
Precision	83.07	91.61	87.51	80.61	86.06	85.28

False Positive FP	19.53	9.2	15.1	20.87	6.6	7.27
False Negative FN	13.47	13.07	64.1	26.6	72.33	71.07
Classify error	0.130	0.087	0.207	0.186	0.310	0.308

Table 4—7 Ensemble Model results

Ensemble models such as Gradient Boosted Trees and the three-classifier model using Logistic Regression, SVM and Neural Net were the models with the higher performance value measures. In comparison, the individual classifier using Logistic Regression outperformed the ensemble models in all the performance measures. In the case of Decision Tree and Naïve Bayes the ensemble models with both these algorithms fared better. Details of the performance measure values for each of the ensemble models that were included in this study is provided below.

This section will discuss the performance of the ensemble models that were used in this study. Bagging and Boosting will be covered in brief as this has already been discussed in the model performance for the individual classifiers.

Gradient Boosted Trees Performance Measure Values

Gradient Boosted (dataset excluding course grades)

	Weighting	SMOTE	Gradient Boosted	Bootstrap Sampling	Stratified Sampling
AUC	0.966	0.958	0.937	0.932	0.922
Accuracy	0.931	0.910	0.870	0.863	0.844
G-mean	0.930	0.910	0.871	0.857	0.830
Kappa	0.861	0.819	0.739	0.721	0.679

f-measure	0.922	0.899	0.858	0.840	0.810
Specificity	94.31	91.72	86.22	90.03	91.67
Sensitivity	91.66	90.19	88.07	81.67	75.23
Precision	92.79	89.68	83.07	86.81	87.81
False Positive FP	8.07	35.2	19.53	14.33	35.40
False Negative FN	9.4	33.2	13.47	20.66	83.80
Classify error	0.069	0.090	0.130	0.137	0.156

Table 4—8 Gradient Boosted Trees results

Gradient boosted trees models have very high AUC scores for all the models. The interesting aspect of these models is that they also registered high values for the other performance measures namely Accuracy, Geometric Mean and f-Measure with Kappa scores that suggest moderate to substantial agreement. The models using weighting and SMOTE had the best performance values across all the measures.

The Gradient Boosted Trees models had high values for specificity, sensitivity/recall and precision. The errors were also minimal as the values were low for classification errors, false positives and false negatives especially in the models using weighting and the original model.

Random Forest Performance Measure Values

Random Forest (dataset excluding course grades)

	Bootstrap Sampling	**Stratified Sampling**	**SMOTE**	**Random Forest**
AUC	0.796	0.780	0.775	0.689
Accuracy	0.708	0.716	0.711	0.692
G-mean	0.627	0.632	0.632	0.592
Kappa	0.376	0.395	0.391	0.338
f-measure	0.519	0.568	0.566	0.516

Specificity	93.74	95.01	95.06	94.87
Sensitivity	41.91	42.07	42.07	37
Precision	84.26	87.87	87	85.28
False Positive FP	8.87	9.9	10.6	7.27
False Negative FN	65.53	98	98	71.07
Classify error	0.292	0.284	0.289	0.308

Table 4—9 Random Forest results

Random Forest does not recognize weights and hence the model with weighting has not been included. Random Forest models had a fair performance in terms of the performance measure values with low kappa scores that show a slight to fair agreement. f-Measure values were also lower indicating low values for precision and recall for these models. Specificity values were high for all the Random Forest models, but recall was low and hence the reason for the low f-measure values. The models achieved moderate precision, but the classification errors and false positives and false negative values were relatively high.

Ensemble models using Vote with 2 classifiers

Dataset excluding courses

	Decision Tree and Naive Bayes	**Decision Tree and Random Forest**
AUC	0.717	0.622
Accuracy	0.690	0.691
G-mean	0.584	0.589
Kappa	0.332	0.334
f-measure	0.506	0.51
Specificity	95.34	95.01

Sensitivity	35.87	36.46
Precision	86.06	85.61
False Positive FP	6.6	10.6
False Negative FN	72.33	107.5
Classify error	0.310	0.309

Table 4—10 Ensemble Model with two Classifiers results

Using the Vote nested operator in Rapid Miner® two ensemble models were built with two classifiers each. Decision tree and Naïve Bayes was one of the models while the other one had Decision tree and Random Forest as the classifiers. As the objective of an ensemble classifier is to improve the performance of the weak individual classifiers. The models that had lower performance values were combined using the Vote operator to measure if model performance would actually improve. All parameters were retained the same as when the classifiers were used individually for model building.

The Decision tree with Naïve Bayes model had better values for AUC, and almost similar values as the Decision Tree and Random Forest model for the other measures such as accuracy, geometric mean, kappa scores and f-measure. It was also observed that there was no significant improvement over the individual classifier performances, but the AUC values mirrored that of Naïve Bayes which was the better performer of the two. Specificity values were high for both models, but sensitivity/recall values suffered indicating that the true positive values were low Precision values were moderately high, and the models had low number of false positive errors**.** Low recall values mean that there is a high rate of false negative errors.

Ensemble models using Vote with 3 classifiers

Using the Vote nested operator in Rapid Miner® three ensemble models were built with three classifiers each. The classifiers in each model were as follows:

1. Decision tree with Neural Net and SVM
2. Decision tree with Naïve Bayes and Neural Net
3. Logistic Regression SVM, Neural Net and SVM

Dataset excluding courses

	Logistic Regression, Neural Net and SVM	Decision Tree, Neural Net and SVM	Logistic Regression, Neural Net and SVM sampling stratified	Decision Tree, Naïve Bayes and Neural Net
AUC	0.917	0.885	0.826	0.743
Accuracy	0.913	0.793	0.814	0.717
G-mean	0.909	0.760	0.807	0.693
Kappa	0.822	0.566	0.62	0.415
f-measure	0.900	0.721	0.784	0.645
Specificity	93.41	92.89	85.27	82.32
Sensitivity	88.41	62.1	76.42	58.35
Precision	91.61	87.51	80.61	72.76
False Positive FP	9.2	15.1	20.87	25.07
False Negative FN	13.07	64.1	26.6	47
Classify error	0.087	0.207	0.186	0.283

Table 4—11 Ensemble Model with three classifiers results

The ensemble model with three classifiers namely Logistic Regression, neural net and SVM had the highest values across all the performance measures. Although the individual classifiers such as Logistic Regression and SVM had good performance measure values the improvement of the Neural Net performance measures as a combined model is evident. This model with stratified sampling also had a consistent good performance across the performance measures with kappa scores suggesting moderate agreement. The model with Decision tree, Naïve Bayes and Neural Net had the lowest performance with moderate values for AUC, accuracy, geometric and f-measure. Kappa scores were low and suggest a slight agreement.

The ensemble model with three classifiers namely Logistic Regression, neural net and SVM had the highest values for specificity thus indicating high values for true negatives. Sensitivity and precision values were also the highest in this model which indicates high values for true positives and low values for false positives. The errors were also lowest in this model. Although the ensemble with Decision Tree, Neural Net and SVM had comparable values for AUC and precision. The recall values were low and correspondingly the false negatives were high.

Bagging and Boosting

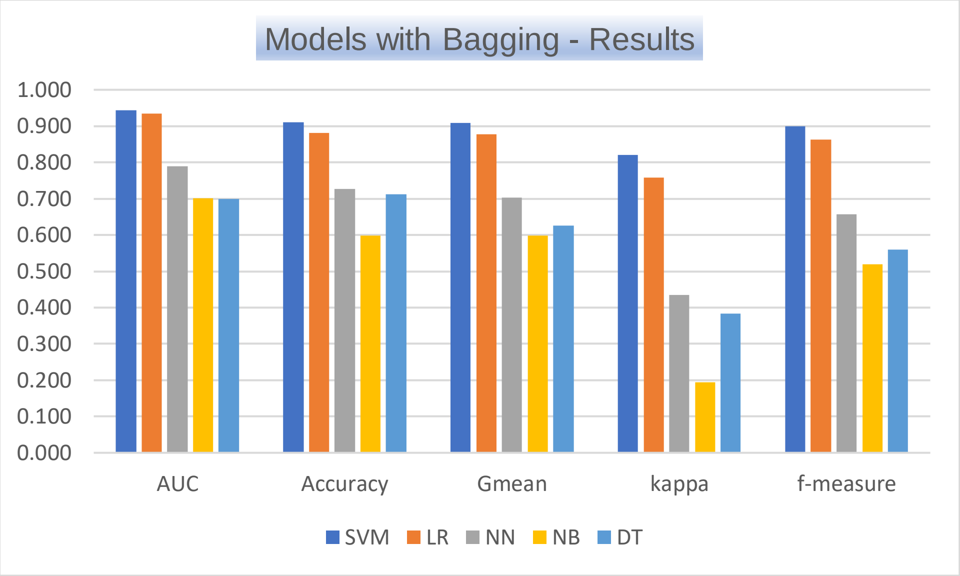


Figure 4-1 Models with Bagging results

SVM and LR had consistent high scores across all the measures with the use of bagging the ensemble operator.

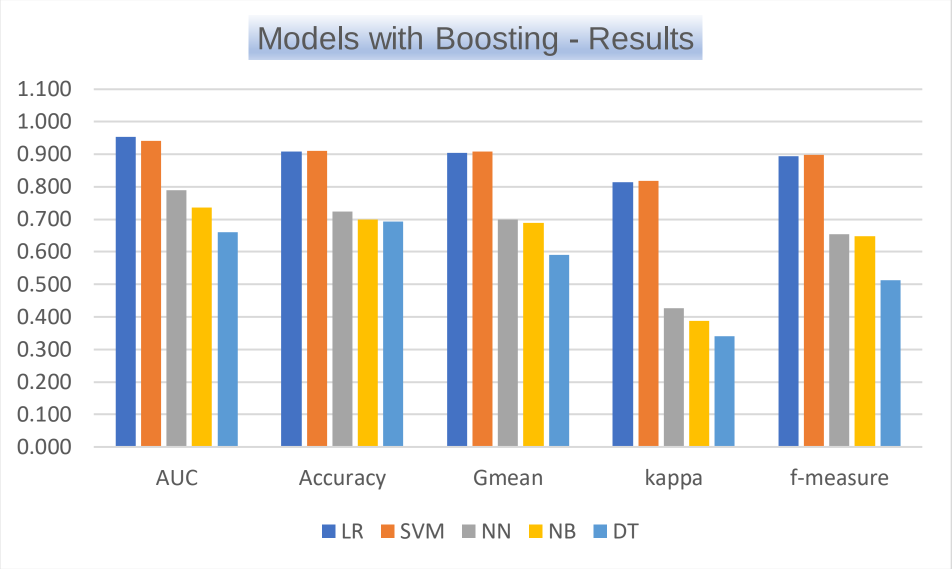


Figure 4-2 Models with Boosting results

LR and SVM once again have the better performance scores with the models using boosting as the ensemble operator. Decision Tree and the Naïve Bayes algorithms had the weaker performance for both bagging and boosting models.

Model performance with Sampling and Weighting

Sampling is used with all the models except for the ensemble models with two and three classifiers using the Vote operator. The sampling methods used in this study are Stratified. Bootstrap and SMOTE. Stratified sampling builds random subsets from the example set ensuring that in a binominal classification each of the subsets contains approximately the same proportions of the two class labels. Bootstrap sampling uses sampling with replacement and hence the samples may not all have unique examples. Synthetic Minority

Oversampling technique (SMOTE) (Chawla *et al.*, 2002) filters the minority class and uses k-nearest neighbors to create synthetic examples that are similar to random examples that are chosen from the whole example set. Weighting was another technique that was employed to study if it had any impact on the model performance. Weights are distributed over the samples such that the weight sums up equally for each label. The answer to the research question 'Does the use of sampling and weighting techniques improve model performance?' is elaborated in this section.

Sampling did improve the performance of the models with algorithms such as Gradient Boosted trees, Logistic Regression and SVM with SMOTE. Stratified and bootstrap sampling also had fairly high-performance measure values with these algorithms. Weighting had a significant impact on the model performance for Gradient Boosted trees.

The results for the various performance measures for each of these models, original models and models with sampling and weighting is available in Appendix 2. The results of the performance measures for each of the models is provided below.

Original Model

The original models are the models for the various algorithms without the use of sampling or weighting techniques.

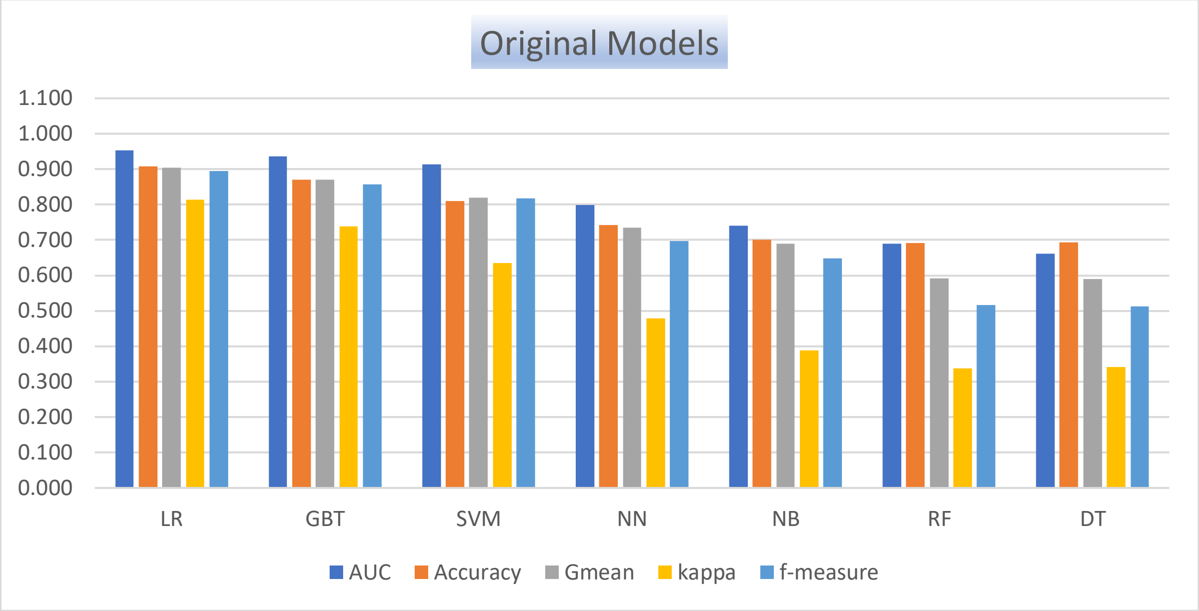


Figure 4-3 Original Model results

Logistic Regression is the model that outperforms all other models in the performance measure values with kappa scores indicating substantial agreement. The Gradient Boosted trees model also had high values for the performance measures, but the kappa scores only showed moderate agreement. Overall the Logistic Regression and Gradient Boosted trees had the better performance when compare to the other models such as SVM, Neural Net, Naïve Bayes, Random Forest and Decision Tree

Sampling

SMOTE

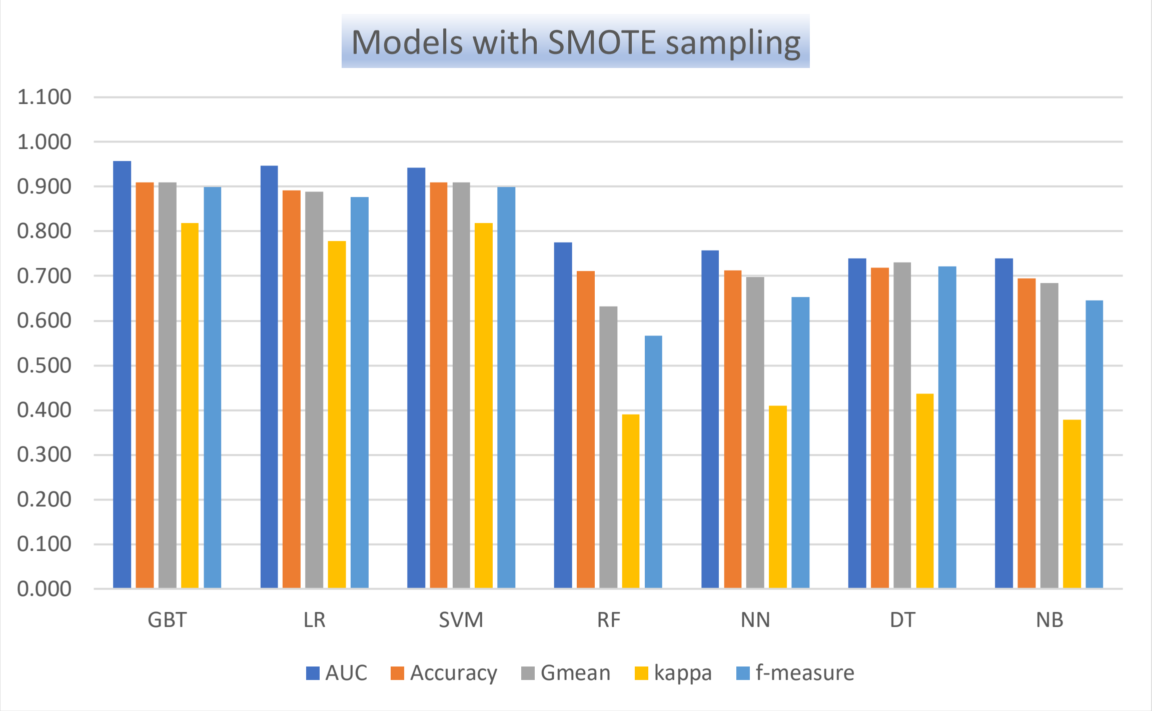


Figure 4-4 Models with SMOTE sampling results

Gradient Boosted Trees, Logistic Regression and SVM models achieved high performance measure values for all the measures in these three models. The kappa scores showed moderate to substantial agreement.

There is a significant drop in performance measure values in the other models such as Random Forest, Neural Net, Decision Tree and Naïve Bayes. The kappa scores for these models also dropped indicating a slight to fair agreement.

Bootstrap Sampling

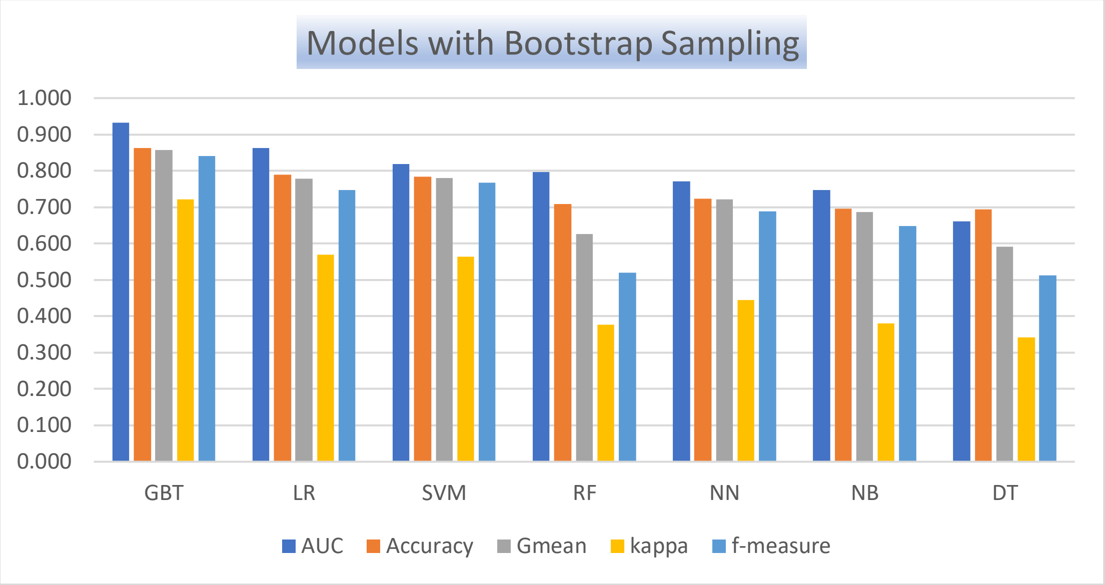


Figure 4-5 Models with Bootstrap Sampling results

Gradient Boosted trees with Bootstrap sampling recorded high AUC values and moderately high values for the other performance measures too indicating that the model performance was high. There is a significant drop in kappa scores in the other models and values of performance measures such as accuracy, geometric mean and f-measure also dropped significantly in the Random Forest, Neural Net, Naïve Bayes and Decision Tree models using SMOTE sampling

Stratified Sampling

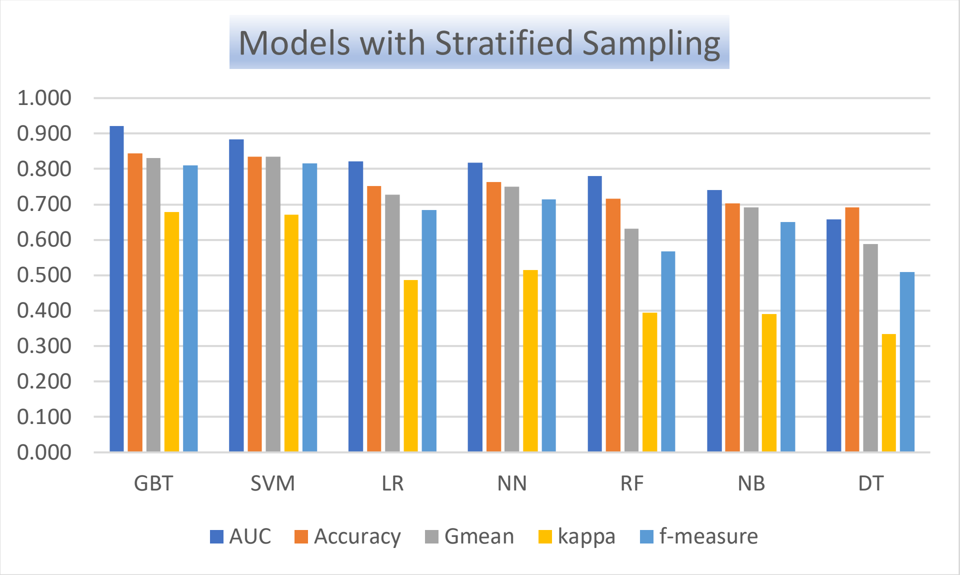


Figure 4-6 Models with Stratified Sampling results

Although the Gradient Boosted Trees had a high AUC value the SVM model with stratified sampling had a slightly lower and preferred AUC value with all other performance measure values comparable to the Gradient Boosted Tree model. Values for accuracy, geometric mean, f-measure and kappa scores dropped for the other models using stratified sampling.

Weighting

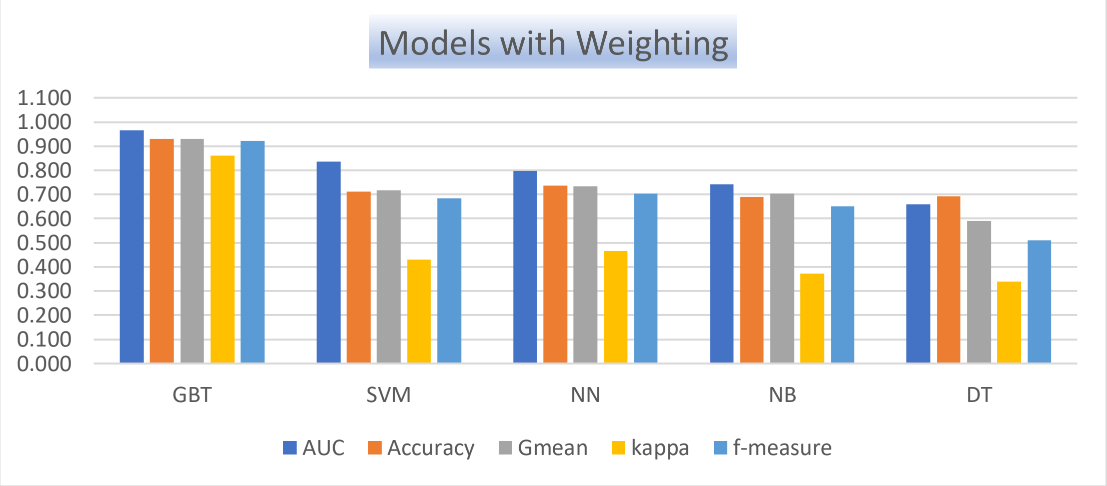


Figure 4-7 Models with Weighting results

Gradient boosted trees had the highest performance values for all measures when compared to the other models that use weighting. Although a significantly high AUC is not considered favorably the high values this model also obtains for other performance measures suggests that it is one of the better performing models when using weighting.

Feature Set selection

Feature selection details for dataset 1 and dataset 2

Principal component analysis (PCA) was conducted on this reduced dataset, using Rapid Miner® to determine which attributes are relevant to this dataset. It is also possible to use PCA for dimensionality reduction. It is recommended that PCA should be evaluated in the context of the data because it has a tendency to give a lot of significance to the noisiest variables. (Kotu and Deshpande, 2014). In this study both PCA and feature weights was employed to highlight the relevant predictors in this dataset. The research question addressed in this section is 'What are the highly relevant predictors in the given dataset?'

The PCA eigen vectors and the eigen values were analyzed for the datasets to determine the most relevant features. While for feature selection by weights the attribute were sorted from largest to smallest and attributes with the higher weight values were considered as relevant. Details of this analysis are provided in the section below.

Dataset 1

Principal Component Analysis

Principal Component Analysis Dataset 1 - Eigen Values			
Component	**Standard Deviation**	**Proportion of Variance**	**Cumulative Variance**
PC 1	1.538	0.263	0.263
PC 2	1.422	0.225	0.487
PC 3	1.033	0.118	0.606
PC 4	0.972	0.105	0.711
PC 5	0.924	0.095	0.806
PC 6	0.815	0.074	0.88
PC 7	0.774	0.067	0.946
PC 8	0.553	0.034	0.98
PC 9	0.421	0.02	1

Table 4—12 Principal Component Analysis Eigen Values Dataset 1

The eigen values show that PC1 to PC7 determine almost 95% of the variance. If the variance threshold is set to 95% then PC1 to PC7 will be considered as most relevant. The eigen vector table is then used to determine the features that PC1 to PC7 relate to, using the highest value in the column.

Eigen Vectors Dataset 1									
Attribute	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
Gender Code	-0.032	0.362	-0.1	0.165	-0.883	0.033	0.218	0.006	-0.037
Campus Code	-0.102	0.148	-0.648	0.656	0.2	-0.009	-0.261	0.013	-0.095
CGPA	-0.427	-0.085	0.204	0.107	-0.09	-0.847	-0.148	-0.091	-0.042
High School Average	-0.572	0.178	-0.116	-0.155	0.049	0.184	-0.021	-0.005	0.753

High School ENGL	-0.465	-0.341	-0.134	-0.031	0.003	0.113	0.385	0.625	-0.312
High School MATH	-0.468	0.348	0.031	-0.257	0.119	0.252	-0.079	-0.441	-0.561
CEPA	-0.07	-0.583	-0.213	0.127	-0.091	0.061	0.413	-0.636	0.077
IELTS Band	-0.14	-0.474	0.125	0.032	-0.36	0.314	-0.713	0.028	-0.041
Absences	0.142	-0.074	-0.66	-0.65	-0.132	-0.262	-0.169	0.022	-0.042

Table 4—13 Principal Component Analysis Eigen Vectors

Based on the highest values (absolute) from the eigen vector matrix it can be concluded that High School Average, CEPA and Absences are highly relevant. Other relevant features include IELTS band, CGPA, and Gender.

Feature Selection by Weights

Feature selection by weights were also analyzed as statistical measures to define the importance of the feature sets. Correlation, Gini Index, Information Gain and PCA were the measures used for this analysis. Feature selection was done using the Feature weight operators in Rapid Miner®. Feature weights by correlation, Gini index, information ratio and PCA were evaluated.

Weight by correlation calculates the correlation between the attributes in the example set and the target variable or label. The higher the weight of the attribute the more relevant it is.

Weight by Gini Index computes the Gini index of the class distribution and uses this to calculate the weight of the attribute with respect to the label attribute. Higher the value more the relevance of the attribute

Weight by Information ratio calculates the weights of the attributes based on the information gain ratio. Higher the value more relevant is the attribute.

Weight by PCA creates weights for the attributes based on a component created by the PCA. Higher values are considered more relevant.

Feature Selection by Weights				
Correlation			**Gini Index**	
Attribute	**Weight**		**Attribute**	**Weight**
High School Average	0.132		High School Average	1
High School ENGL	0.086		High School MATH	0.654
High School MATH	0.057		High School ENGL	0.521
Absences	0.024		IELTS Band	0.153
IELTS Band	0.018		Absences	0.149
CEPA	0.012		CEPA	0.138
Campus Code	0.003		Campus Code	0.042
Gender Code	0		Gender Code	0
Information Ratio			**PCA**	
Attribute	**Weight**		**Attribute**	**Weight**
High School Average	0.160		CEPA	0.241
CEPA	0.136		Absences	0.141
IELTS Band	0.124		IELTS Band	0.135
High School ENGL	0.111		Campus Code	-0.192
High School MATH	0.107		Gender Code	-0.217
Absences	0.087		High School ENGL	-0.237
Campus Code	0.008		High School Average	-0.618
Gender Code	0		High School MATH	-0.618

Table 4—14 Feature selection by Weights Dataset 1

High School Average ranks the highest in more than one feature selection model. The other attributes that also have a high ranking include CEPA and Absences. Other features that got a high rank were High School English and Math and the IELTS band.

Dataset 2

Principal Component Analysis

Principal Component Analysis Dataset 2- Eigen Values

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	2.738	0.416	0.416
PC 2	1.718	0.164	0.58
PC 3	1.435	0.114	0.695
PC 4	1.176	0.077	0.772
PC 5	0.919	0.047	0.819
PC 6	0.907	0.046	0.864
PC 7	0.759	0.032	0.896
PC 8	0.68	0.026	0.922
PC 9	0.599	0.02	0.942
PC 10	0.532	0.016	0.958
PC 11	0.41	0.009	0.967
PC 12	0.376	0.008	0.975
PC 13	0.313	0.005	0.98
PC 14	0.298	0.005	0.985
PC 15	0.292	0.005	0.99
PC 16	0.262	0.004	0.994
PC 17	0.243	0.003	0.997
PC 18	0.231	0.003	1

Table 4—15 Principal Component Analysis Eigen Values Dataset 2

Principal Component Analysis Dataset 2 - Eigen Vectors

Attribute	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CEPA	0.1	-0.03	0.166	0.611	0.33	0.2	0.452	0.472	-0.061	0.078
IELTS	0.035	0.004	0.212	0.497	-0.667	0.373	-0.324	-0.103	0.037	0.008
C1_Project	-0.127	-0.25	0.308	0.126	-0.229	-0.563	0.277	-0.055	0.594	0.042
High School Average	-0.134	-0.144	-0.144	-0.408	-0.477	0.364	0.612	0.177	-0.037	0.04
C1_Research Project	-0.142	-0.295	0.33	-0.007	-0.183	-0.398	-0.001	0.099	-0.747	-0.004
C1_Quiz2	-0.16	-0.271	0.254	0.076	0.284	0.301	0.265	-0.753	-0.081	-0.045
C1_Quiz1	-0.172	-0.284	0.32	-0.195	0.139	0.212	-0.235	0.326	0.169	-0.664
C1 – FE	-0.185	-0.288	0.27	-0.219	0.158	0.222	-0.301	0.202	0.104	0.698

C3_Project	-0.251	-0.179	-0.369	0.207	-0.007	-0.092	-0.046	-0.017	-0.088	-0.012
C3_Practical	-0.263	-0.235	-0.322	0.152	-0.02	0.027	-0.059	0.024	0.043	-0.106
C3_Quiz 1	-0.264	-0.214	-0.34	0.154	0.031	-0.036	-0.045	0.011	-0.029	0.022
C3_FE	-0.273	-0.251	-0.256	0.102	0.052	0.021	-0.101	0.051	0.096	0.049
C2_Pract 1	-0.291	0.31	0.03	0.026	-0.004	-0.073	0.055	0.053	0.01	0.035
C2_Pract2	-0.301	0.272	0.067	0.028	0.015	-0.035	-0.004	-0.031	-0.003	0.055
C2_Project	-0.305	0.27	0.086	0.028	-0.036	-0.037	0.013	0.018	-0.048	0.074
C2_Quiz2	-0.312	0.226	0.103	0.036	-0.004	0.054	0.044	-0.001	-0.096	-0.138
C2_Quiz1	-0.315	0.242	0.089	-0.008	0.042	0.022	0.011	0.017	0.027	-0.102
C2-FE	-0.323	0.205	0.102	-0.042	0.033	0.078	0.004	0.014	0.069	0.085

Table 4—16 Principal Component Analysis Eigen Vectors Dataset 2

The eigen values of the PCA in dataset 2 indicate that PC1 to PC10 make up 95% of the variance. Hence the features associated with PC1 to PC10 with the higher eigen vector values will be considered as relevant for this dataset.

In dataset 2, using the eigen values and eigen vectors from the PCA, it can be concluded that CEPA scores, High School Average and IELTS are relevant features. In addition, the final exam, project assessment and quiz 2 are some of the other features that are selected. It should be noted that none of the Course 2 features are selected and yet practically it will not be feasible to ignore the performance of students in this course or other assessments that are not deemed relevant by the PCA.

Feature Selection by Weights

Feature Selection by Weights - Dataset 2

Correlation			**Gini Index**	
Attribute	**Weight**		**Attribute**	**Weight**
C1 – FE	1		C3_FE	1
C1_Quiz1	0.975		C1 - FE	0.977
C3_FE	0.689		C1_Quiz1	0.801
C1_Quiz2	0.675		C2-FE	0.631
High School Average	0.533		C3_Practical	0.620
C3_Practical	0.524		C3_Quiz 1	0.502
C1_Research Project	0.495		C1_Quiz2	0.404

C2-FE	0.458
C3_Quiz 1	0.429
C1_Project	0.364
C2_Quiz1	0.308
C2_Quiz2	0.282
C3_Project Output	0.195
C2_Project	0.14
C2_Pract2	0.122
CEPA	0.09
IELTS	0.049
C2_Pract 1	0

C2_Quiz1	0.403
High School Average	0.342
C2_Project	0.284
C1_Project	0.274
C1_Research Project	0.236
C2_Quiz2	0.201
C3_Project Output	0.172
C2_Pract2	0.121
C2_Pract 1	0.097
IELTS	0.004
CEPA	0.000

PCA	
Attribute	**Weight**
C2-FE	1
C2_Quiz1	0.971
C2_Quiz2	0.960
C2_Project	0.938
C2_Pract2	0.922
C2_Pract 1	0.888
C3_FE 30%	0.826
C3_Quiz 1	0.793
C3_Practical	0.792
C3_Project Output	0.751
C1 – FE	0.522
C1_Quiz1	0.475
C1_Quiz2	0.433
C1_Research Project	0.373
High School Average	0.343
C1_Project	0.319
CEPA	0.226
IELTS	0

Information Gain Ratio	
Attribute	**Weight**
C3_FE	1
C1 - FE	0.826
C2-FE	0.665
C1_Quiz1	0.602
C3_Practical	0.572
C2_Quiz1	0.492
C2_Project	0.385
C3_Quiz 1	0.384
C2_Quiz2	0.321
High School Average	0.307
C1_Quiz2	0.294
C1_Research Project	0.230
C1_Project	0.109
IELTS	0.086
C3_Project	0.029
C2_Pract 1	0.025
C2_Pract2	0.025
CEPA	0

Table 4—17 Feature selection by Weights Dataset 2

Features selection by weights suggests that course final exam grade is highly relevant as it does have the highest weight across the various measures. As the three courses are mandatory for all Year 1 students', it can also be concluded from a domain perspective that final exam scores in these courses is a factor that influences the target variable namely if the student is ‘At Risk’ or in ‘Good Standing’. Quizzes in Course 1 and Course 2 also have recorded high weights in the weight by correlation and weight by PCA measures.

CEPA and IELTS scores are in the lower range of relevance in almost all the techniques used in feature selection by weight. Ranking of attributes is similar with weight by information gain ratio and Gini index. High School average had a mid-range value in weight by correlation but had similar weight values in all the other measures such as Information Gain, PCA and Gini index weight values.

Early Prediction of At Risk

Based on the performance evaluation of the classification models for the first research question, models with the higher values for AUC, f-measure, accuracy, geometric mean and kappa scores are selected to address the third research question: “How early can we predict student ‘At Risk’ status in terms of the course work done in Semester 1 of the program?”

Methodology - Early Prediction of At Risk

Coursework and final exam grades of about 400 students across over ten different campuses were processed for three different mandatory courses in semester 1 of the program. Details of the assessments are given in the table below:

Exam Type	Completed	Weighting
Quiz 1 - Coursework	Week 3	10%
Quiz 2 - Coursework	Week 6	15%

Practical Exam - Coursework	Week 10	15%
Project	Week 15	30%
Final Exam	Week 17	30%

Table 4—18 Early Prediction Assessment details

Based similar work done in other studies in this area, the models were selected using the assessment scores to determine the performance measures for prediction at different stages of the semester. (Lehr *et al.*, 2016; Adejo and Connolly, 2018; Burgos *et al.*, 2018).

The dataset consisted of 18 attributes and 435 records. The models were built as follows:

Model 1 – Quiz 1 from all three courses, High School Average, CEPA score, IELTS band, GPA rating (label)

Model 2- Model 1 attributes and Quiz 2 grades from all courses

Model 3 - Model 2 attributes and Practical assessment grades from all course

Model 4 – Model 3 attributes and Project grade for all course

Model 5 – Model 4 attributes and Final Exam grades for all courses

The algorithms used on these models were the ones that had a good performance with Dataset 1. (Results). The results for the various models using the performance measures is available in Appendix 3.

The best performing model at the various stages in the semester is Gradient Boosted Trees with weighting. It consistently outperformed all other algorithms used. A summary of the scores achieved is provided below:

Gradient Boosted Trees with Weighing Models for Early At-Risk Prediction

	AUC	GM	f-measure	Accuracy	kappa	Precision	Recall
Quiz 1	0.868	0.803	0.832	0.802	0.592	0.878	0.796
Quiz 1 and 2	0.893	0.807	0.825	0.800	0.594	0.893	0.774
Quizzes and Practical	0.903	0.808	0.852	0.818	0.616	0.863	0.848
Quizzes Practical and Project	0.911	0.832	0.863	0.835	0.653	0.892	0.841
Coursework and Final Exams	0.923	0.840	0.880	0.851	0.682	0.88	0.88

Table 4—19 Early Prediction of At-Risk Model results

.

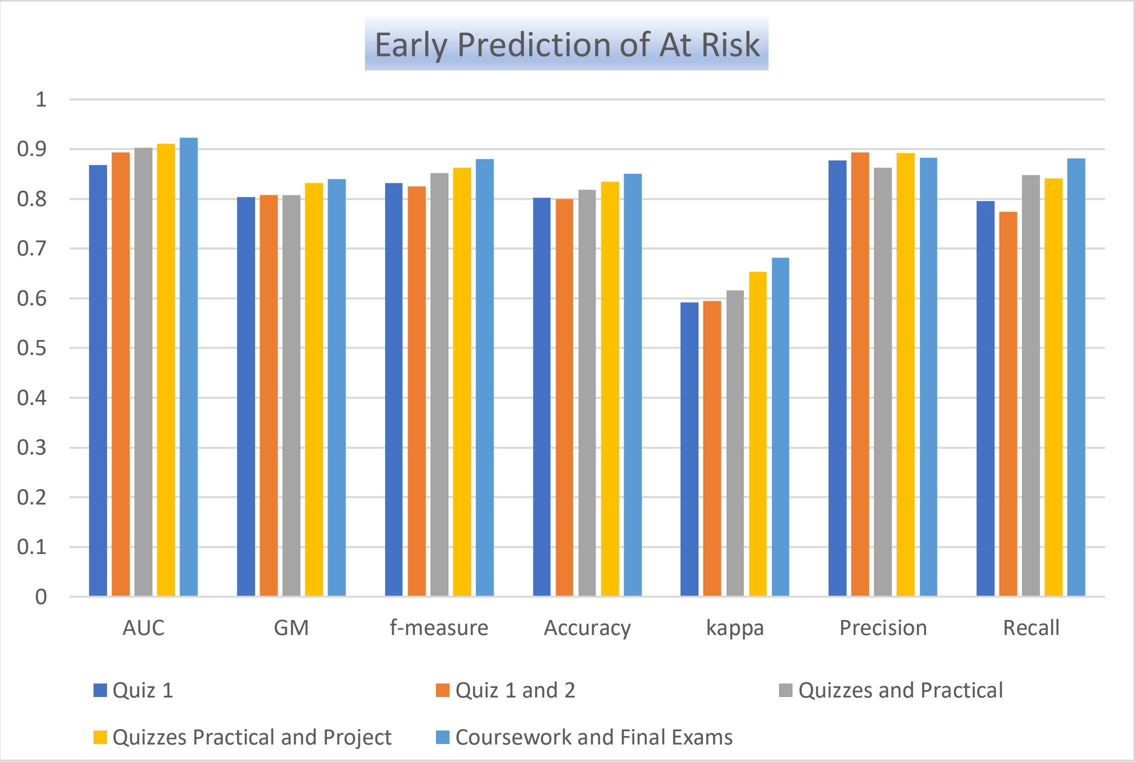


Figure 4-8 Early Prediction of At-Risk GBT results

Predictably the model that includes all the assessment scores (coursework and final exam) is the one with the highest AUC and other performance measure values. The variance in the performance measures values is minimal between the last three assessment stages. The AUC ranges from 0.893 for both the quizzes to about 0.923 for all the assessments. The difference in values for the other performance measures for these models follow a similar pattern. Precision and Recall improve with each stage of the model assessment. Although the model which included the coursework and final exam grades had the highest values for the performance measures the focus on early prediction would require that it is done earlier in the semester and not after students have finished their final exam. In this context, early prediction is possible after the students have completed Quiz 1 and Quiz 2 in all the courses. This would be weighted at about 25% of their final grade in a course and completed about half way through the semester. Hence the timing too would be effective in terms of initiating interventions to improve academic performance.

Chapter 5 Discussion

CRISP DM

The CRISP-DM framework has provided a structured approach to the various stages in this research. Business and data understanding required knowledge gained from the literature review and a thorough understanding of the business rules and context of this study. It helped to frame the research questions that defines the purpose and goal of conducting this study. Data preparation demanded a thorough and detailed approach to ensure data quality for effective modeling. It provided the opportunity of reviewing and remembering all the processes that was covered in the data mining course to get the datasets ready for modeling. The final stage of the CRISP-DM framework is knowledge that is meant to describe how business knowledge can be gained by using data mining to understand patterns and successfully transform a business issue into a data problem. Mining the student performance data provided an insight into how pre college and course performance can influence student status in the freshmen year, and this could add great value to the business in arresting attrition and motivating students to perform better.

Data Mining Solution

The primary objective of this research was to determine the best approach to identify students who are ‘At Risk’ of failing by using pre college and other academic data. This provided the impetus to expand this research study to evaluate classification models using multiple algorithms, explore the use of sampling and weighting to address data imbalance and improve model performance, determine how performance is impacted by using ensemble models, assess the importance of features in datasets and use of the data on course and final exam assessments to compare accuracies of early risk prediction. As model evaluation is the main focus it warranted a structured approach to employing the relevant performance measures based on current research in performance prediction. Literature review of the current work on performance prediction provided key insights into the methodology used for performance prediction. The process of reviewing current studies on the topic of At Risk and student performance using educational data mining resulted in reading and learning from over 70 papers in addition to books and articles on this and related topics. The immense knowledge and direction it provided was instrumental in the approach and methodology used to complete this study.

Classification Models

In order to achieve the dissertation objectives, it is concluded from the literature review that a majority of the studies used classification models with multiple algorithms. Studies similar to the one undertaken in this research provide an insight into how classification as a data mining approach can be used to predict ‘At Risk’ status for students in their freshmen or first year at University. A large number of researchers employed the use of multiple models with various algorithm and a similar approach is adopted in this study. The algorithms were chosen based on their usage and performance in similar studies. Decision Tree, Naïve Bayes, Neural Net, SVM and Logistic Regression were the algorithms chosen. Although some researchers have reported a better performance with Ensemble classifiers this technique has **not** been explored by many studies as part of the data mining models used in prediction performance. This is considered as a gap that needed to be addressed. Ensemble models are used to combine the effects of the weaker base classifiers and produce a model with higher accuracy in performance measures. Hence ensemble classifiers using boosting, bagging, the vote operator Gradient Boosted Trees and Random Forest was included in this research. A significant majority of the literature reviewed in ‘At Risk’ prediction do not address the issue of data imbalance, this area is relatively new to researchers according to (Thammasiri *et al.*, 2014). To address this gap models in this study were run with recommended imbalance techniques using sampling and weighting. Sampling and weighting were used to ascertain whether it helps to improve model performance and address the issue of data imbalance. A comprehensive study of the best performing models required that these techniques were used in the models to analyze its impact on model performance. Sampling techniques such as Stratified sampling, Bootstrap sampling and SMOTE sampling has been used with all the models. Weighting has also been implemented with the models that recognize weights. Logistic Regression and Random Forest models ignore weights and hence weighting is not used with these algorithms.

Performance Measures

Evaluating model performance is usually restricted to the confusion matrix and accuracy readings and this was the case in some of the literature that was reviewed. Learning from other researchers the importance of using multiple performance measures contributed immensely to the performance evaluation process in this study. Performance measures used in different studies is collectively employed in this study thereby negating the disadvantages of using one or two performance measures. The models are evaluated based on its efficacy to perform well in more than one measure which is a more robust method of performance evaluation. Accuracy is a common performance measure that was used in most of the literature reviewed for prediction performance in education data mining. As data imbalance was being considered it was important to evaluate using other measure besides accuracy. In addition to accuracy, the use of geometric mean and specificity are recommended as these are not sensitive to imbalanced datasets (Tharwat, 2018) . Other measures recommended are AUC, f-measure and kappa scores. Since a majority of the studies recommend the use of multiple performance measures the models were evaluated using measures such as AUC, geometric mean (GM), accuracy, f-measure and kappa scores. The advantage of this approach is that even when the AUC value seems inflated model performance is judged based on the other performance measure values. Model errors were analyzed using the values of classify errors, false positives and false negatives. Precision and recall values were also studied to rate the accuracy of the model predictions.

Model Performance

The best models in this study achieved AUC values above 0.85 and high values in the other performance measures too. Logistic Regression is the only individual classifier model that achieved the highest AUC value without the use of sampling or weighting. The SVM models using bagging and boosting and the SVM model with SMOTE sampling had marginally better values for the other measures such as accuracy, geometric mean, f-measure and kappa scores when compared to the Logistic Regression model. Errors were the lowest in these models while precision and recall was high. The use of stratified sampling also improved the performance for Neural Net models, but error values were high, and precision and recall were low. Naïve Bayes with bootstrap and stratified sampling performed better than the Decision Tree model but performance measure values were lower than 0.8, errors were high, and precision and recall had a poor performance too. Decision tree model performance did not necessarily improve with sampling, weighting or ensemble model operators such as bagging and boosting. The performance measure values were comparable between the original model, models using sampling and weighting and the Ensemble models.

Gradient boosted trees and the three-classifier model with the Vote operator that included the algorithms of Logistic Regression, Neural Net and SVM were the better performing Ensemble models. Gradient boosted trees had a better performance with sampling and weighting and recorded the highest values across all measures. Precision and recall were high in both models but the Gradient Boosted trees with weighting recorded the lower value in false positive and false negative errors. The tendency of over fitting which is prevalent in Gradient Boosted trees was addressed by using a 15-fold cross validation using stratified sampling with this model. The model not only achieved high values with AUC but also had significantly high values for f-measure, geometric mean (GM), accuracy and kappa scores that indicated an almost perfect agreement. The three-classifier model also performed well with low errors and high precision and recall.

It can thus be concluded the best models with high performance measures values, good precision and recall and low errors are as follows:

- Gradient Boosted trees with Weighting

- SVM with Bagging and SVM with Boosting
- SVM with SMOTE
- Logistic Regression
- Ensemble with LR, SVM and NN

Ensemble operators and sampling and weighting techniques clearly help in improving model performance.

Feature sets

As model performance was crucial to this study it had to be supported by using feature selection techniques which is a significant contributor to how a model performs. The process of feature set selection was restricted to identifying the key predictors in the dataset. Due to privacy and time constraints in securing the necessary approvals only the attributes that did not disclose private or social information about individual students and is relevant to the study is included as part of the dataset.

Principal component analysis and feature set weights were used to determine the relevant predictors. In PCA the eigen values were used to identify the attributes that contributed to a 95% variance threshold. The eigen vector matrix was then used to determine the features that contribute to these principal components. Feature selection by weights is done using the statistical measures of information gain, correlation, PCA and Gini index. The top absolute weights for each measure was then used to determine the relevant features.

In Dataset 1 which is used for performance prediction models the eigen values of the principal component analysis shows that 95% of the variance is determined by PC1 to PC7. Using the eigen vector matrix the associate features for PC1 to PC7 are High School Average, CEPA scores, Absences, Gender, Campus, CGPA and IELTS band. The feature selection by weights also highlights that high school average is a key feature. Absences, CEPA scores and High School Math and English scores are also seen as relevant predictors. The Decision Tree algorithm had a poor model performance but information from the tree provides a valuable insight into the classification process. This information is not available in other models such as SVM, Naïve Bayes or Neural Net. A description of the trees for the algorithms Decision Tree, Gradient Boosted Trees and Random Forest is provided in Appendix 4. Based on this it can be seen that all these algorithms also have High School Average as a key feature followed by CEPA. The other selected features are Absences, High School English and High School Math.

It can be this concluded that they key predictors for this dataset are High School Average, CEPA scores, Absences followed by High School English and High School Math. Gender and campus can also be considered as fairly relevant as it was selected by more than one measure for feature selection by weights and by the Gradient Boosted trees algorithm.

Hence the current system of labelling students as ‘At Risk’ based on their attendance is valid. This could be reinforced by using the High School Average and CEPA scores to determine potentially At-Risk students when they start the freshman year in the program.

In dataset 2, which is used for early prediction of At Risk the eigen values in the principal component analysis PC1 to PC10 made up 95% of the variance. The eigen vector matrix is then used to identify the relevant features. The final exams in the three courses featured as relevant in both the PCA and feature by weights analysis. High School average is relevant in the PCA method but was a mid-performer in the feature selection by weights. CEPA is prominent in the PCA but both CEPA and IELTS scores were low performers in the feature selection by weights.

The course final exam being chosen as highly relevant complies with the domain perspective as the three courses are mandatory for all students and the final exam is weighted at 30% of the final grade. Performance in the first semester could be highly related to their school performance and students who have performed well in school should generally be able to deliver a good performance in the freshman year. Hence High School average is also relevant in accordance with the feature set selection process. Although CEPA and IELTS scores were not significant in the feature selection process it can be argued that a student with a good high school average will also perform well in the CEPA and perhaps the IELTS too.

Feature creation led to the attributes Academic Standing in Dataset 1 and GPA rating in Dataset 2. GPA is recommended by researchers as one of the best predictors of student performance. (Thammasiri *et al.*, 2014). These attributes were used as the labels for binominal classification in this study.

Early Prediction of At Risk

The models that had achieved high performance measure values was used in this part of the study to determine the levels of accuracy that could be achieved using assessment scores of the students. Gradient Boosted trees with weighting is the model that achieved the highest values across all the performance measures in all the models. The best accuracy was predictably the model that used all the coursework grades and the final exam grades. The variance in performance measure values in the model using the performance of the students in Quiz 1 and Quiz 2 is not significant when compared to the models that include more assessments and the best model that includes all the assessments. In the context of early prediction, it can be argued that the model with coursework grades from Quiz 1 and Quiz 2 in all courses has sufficient performance accuracy to allow prediction to happen at this stage. This would mean that 'At Risk' status of students can be identified when they have completed Quiz 1 and Quiz2 of all three courses which has a combined weighting of 25% of the grade in each course. Also, since this would be complete at around Week 6 or so it would provide adequate intervention time for students to be moved out of the 'At Risk' status. If we consider the best model, then prediction will happen at the end of the semester when all the coursework and final grades have been achieved. An intervention at this stage would be futile in terms of betterment of student academic performance.

Chapter 6 Conclusion

Working on this research project has given me a phenomenal experience of successfully completing the process of a research study on performance prediction using a data mining approach. As a passionate educator my initial quest was merely to use data mining to improve the current system of identifying students at risk of failing or dropout. The literature review process allowed me the opportunity to expand this objective and explore the various facets of data mining and how it could be used to achieve the goals of this study.

Using various data mining algorithms and multiple models allowed me the opportunity for an in-depth study of how data mining can be used in performance prediction. The process of determining model efficiency required a good understanding of the various algorithms, their parameters in Rapid Miner® and the use of multiple performance measures. Gradient Boosted trees, an Ensemble model with weighting had the best performance for AUC values while SVM with Bagging and Boosting had better values in all the other performance measures. Conducting this research study using the various algorithm and ensemble models with sampling and weighting clearly established how Ensemble models can achieve better performance and the use of sampling and weighting enhances this performance. The use of multiple performance measures is also fairly unique to this study and is inspired by the work done by other researchers where multiple measures such as AUC, f-Measure, geometric mean, accuracy, kappa scores sensitivity and specificity are used (He and Garcia, 2009; Thammasiri *et al.*, 2014; Tharwat, 2018).

Feature set selection techniques using principal component analysis and feature selection by weights was done with the aim of establishing the most relevant features. High School Average featured in both methods, PCA and feature selection by weights while CEPA and IELTS scores were identified as relevant by PCA. These are some of the entry level criteria for student admission and hence clearly relevant for At Risk prediction. Absences and other high school grades in English and Math were also selected as relevant which complies with the factors that affect student performance from a domain perspective. The best models were used to determine how early At-Risk status can be predicted using the course assessments and final exam grades. Based on the work done by (Burgos *et al.*, 2018) student assessment scores were analyzed by each assessment activity progressively through the semester to resolve how accurately early At Risk status can be predicted. This study established that early risk prediction although most accurate when done at the end of the semester could also be initiated when students have completed coursework assessments that are at least 25% of the weighting of the course grade. PCA and feature selection by weights for this dataset (Dataset 2) identified high school average the course final exam scores as highly relevant. Doing well in the course final exam can actually help improve student performance as the weighting of this assessment is 30% of the course grade, proving that these features are relevant to performance prediction

This study enriched my knowledge and capabilities in data mining and research projects and taught me the intricacies of applying current research to solve a business problem. The significance how research findings can be used to publish and enhance current literature in this field is reinforced while doing this kind a research study.

Limitations

This study is limited to a dataset that contains student information from one program in a higher education university. The study could be extended to multiple programs within the university or even multiple universities. Although the study has been performed only on one program it can be easily replicated for other programs within the University. In addition, since the high-performance classification models have been determined much of the initial work in extending this study has already been completed. The datasets considered in this research is restricted to pre-college and other academic data. No personal or social information of the students is included.

Future Improvements

Feature selection is one of the areas in this study that could be improved further to analyze features of the dataset and gather results on the model performance on various feature sets. PCA and feature weighting has been implemented in this study to gain an idea about relevant features in the dataset. This needs to be further enhanced to study the effects of the use multiple feature sets on the modeling performance and evaluating the predictors of At Risk. Gathering information on additional student attributes related to personal and social factors using a survey could result in a more robust model for prediction.

Class imbalance has been addressed using sampling and weighting techniques. This could be improved by adding a cost benefit analysis to investigate the cost of misclassification. This information will allow a more thorough analysis to select the best performing classification models for this study.

References

Abu Saa, A. (2016) ‘Educational Data Mining & Students’ Performance Prediction’, *International Journal of Advanced Computer Science and Applications*.

Adejo, O. W. and Connolly, T. (2018) ‘Predicting student academic performance using multi-model heterogeneous ensemble approach’, *Journal of Applied Research in Higher Education*, 10(1), pp. 61–75. doi: 10.1108/JARHE-09-2017-0113.

Aggarwal, by C. C. (2015) *Data Classification*. Chapman and Hall/CRC.

Agnihotri, L. and Ott, A. (2014) ‘Building a Student At-Risk Model : An End-to-End Perspective’, *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, pp. 209–212.

Akangah, P. *et al.* (2018) ‘2018 ASEE Southeastern Section Conference Predicting Academic Achievement in Fundamentals of Thermodynamics using Supervised Machine Learning Techniques’.

Amrieh, A., Hamtini, T. and Aljarah, I. (2016) ‘Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods’, 9(8), pp. 119–136.

An, P.-N., Steinbach, M. and Kumar, V. (2013) *Introduction to data mining*. Pearson Education India.

Asif, R. *et al.* (2017) ‘Analyzing undergraduate students’ performance using educational data mining’, *Computers and Education*. Elsevier Ltd, 113, pp. 177–194. doi: 10.1016/j.compedu.2017.05.007.

Aulck, L. *et al.* (2016) ‘Predicting Student Dropout in Higher Education’. doi: 10.1002/prot.24187.

Awad, M. and Khanna, R. (2015) *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress.

Azziaty Binti Abdul Rahman, N., Lam Tan, K. and Kim Lim, C. (2016) ‘Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students’, *Journal of Telecommunication, Electronic and Computer Engineering*, 9(2), pp. 2289–8131.

Baker, R. S. *et al.* (2015) ‘Analyzing Early At-Risk Factors in Higher Education e- Learning Courses’, *Proceedings of the 8th International Conference on Educational Data Mining*, pp. 150–155.

Baker, R. and Yacef, K. (2009) ‘Journal of Educational Data Mining JEDM.’, *JEDM | Journal of Educational Data Mining*, 1(1), pp. 3–17. Available at: http://jedm.educationaldatamining.org/index.php/JEDM/article/view/8.

Ben-David, A. (2008) ‘Comparison of classification accuracy using Cohen’s Weighted Kappa’, *Expert Systems with Applications*. Elsevier, 34(2), pp. 825–832.

Burgos, C. *et al.* (2018) ‘Data mining for modeling students’ performance:

A tutoring action plan to prevent academic dropout', *Computers and Electrical Engineering*, 66, pp. 541–556. doi: 10.1016/j.compeleceng.2017.03.005.

Chawla, N. V *et al.* (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, 16, pp. 321–357.

Conijn, R. and Zaanen, M. (2017) 'Predicting student performance with Neural Networks', (May). Available at: http://arno.uvt.nl/show.cgi?fid=143628.

Costa, E. B. *et al.* (2017) 'Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses', *Computers in Human Behavior*. Elsevier Ltd, 73, pp. 247–256. doi: 10.1016/j.chb.2017.01.047.

Deshpande, Bala ; Kotu, V. (2018) *Data Science, 2nd Edition*. Morgan Kaufmann.

Dutt, A., Ismail, M. A. and Herawan, T. (2017) 'A Systematic Review on Educational Data Mining', *IEEE Access*, 5, pp. 15991–16005. doi: 10.1109/ACCESS.2017.2654247.

Ferri, C., Hernández-Orallo, J. and Modroiu, R. (2008) 'An experimental comparison of performance measures for classification', *Pattern Recognition Letters*. Elsevier B.V., 30(1), pp. 27–38. doi: 10.1016/j.patrec.2008.08.010.

Géron, A. (2017) *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. ' O'Reilly Media, Inc.'

He, H. and Garcia, E. (2009) 'Learning from imbalanced data', *Ieee Transactions on Knowledge and Data Engin*, 21(9), pp. 1263–1284. doi: 10.1109/TKDE.2008.239.

Jayaprakash, S. M. *et al.* (2014) 'Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative', *Journal of Learning Analytics*, 1(1), pp. 6–47. doi: 10.18608/jla.2014.11.3.

Jeni, L. A., Cohn, J. F. and De La Torre, F. (2013) 'Facing imbalanced data - Recommendations for the use of performance metrics', *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*. IEEE, pp. 245–251. doi: 10.1109/ACII.2013.47.

Kabakchieva, D. (2013) 'Predicting Student Performance by Using Data Mining Methods for Classification', *Cybernetics and Information Technologies*, 13(1), pp. 61–72. doi: 10.2478/cait-2013-0006.

Kang, K. and Wang, S. (2018) 'Analyze and Predict Student Dropout from Online Programs', pp. 6–12. doi: 10.1145/3193077.3193090.

Karimi-Alavijeh, F., Jalili, S. and Sadeghi, M. (2016) 'Predicting metabolic syndrome using decision tree and support vector machine methods', *ARYA Atherosclerosis*, 12(3).

Kotsiantis, S., Patriarcheas, K. and Xenos, M. (2010) ‘A combinational incremental ensemble of classifiers as a technique for predicting students’ performance in distance education’, *Knowledge-Based Systems*. Elsevier B.V., 23(6), pp. 529–535. doi: 10.1016/j.knosys.2010.03.010.

Kotu, V. and Deshpande, B. (2014) *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

Krawczyk, B. (2016) ‘Learning from imbalanced data: open challenges and future directions’, *Progress in Artificial Intelligence*, 5(4), pp. 221–232. doi: 10.1007/s13748-016-0094-0.

Kumari, P., Jain, P. K. and Pamula, R. (2018) ‘An efficient use of ensemble methods to predict students academic performance’, *Proceedings of the 4th IEEE International Conference on Recent Advances in Information Technology, RAIT 2018*. IEEE, pp. 1–6. doi: 10.1109/RAIT.2018.8389056.

Lakkaraju, H. *et al.* (2015) ‘A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes’, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’15*, pp. 1909–1918. doi: 10.1145/2783258.2788620.

Larose, D. T. (2015) *Data mining and predictive analytics*. John Wiley & Sons.

Lehr, S. *et al.* (2016) ‘Use educational data mining to predict undergraduate retention’, *Proceedings - IEEE 16th International Conference on Advanced Learning Technologies, ICALT 2016*, (1), pp. 428–430. doi: 10.1109/ICALT.2016.138.

M., R., F., N. and A., A. (2018) ‘Predicting and Analysis of Students’ Academic Performance using Data Mining Techniques’, *International Journal of Computer Applications*. IEEE, 182(32), pp. 1–6. doi: 10.5120/ijca2018918250.

Márquez-Vera, C. *et al.* (2016) ‘Early dropout prediction using data mining: A case study with high school students’, *Expert Systems*, 33(1), pp. 107–124. doi: 10.1111/exsy.12135.

Mhetre, V. and Nagar, M. (2018) ‘Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA’, *Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017*, 2018–Janua(Iccmc), pp. 475–479. doi: 10.1109/ICCMC.2017.8282735.

National Strategy for Higher Education 2030 (no date). Available at: https://www.moe.gov.ae/En/MediaCenter/News/Pages/h2030.aspx (Accessed: 28 September 2017).

Peña-Ayala, A. (2014) ‘Educational data mining: A survey and a data mining-based analysis of recent works’, *Expert Systems with*

Applications, 41(4 PART 1), pp. 1432–1462. doi: 10.1016/j.eswa.2013.08.042.

Putler, D. S. and Krider, R. E. (2015) *Customer and business analytics: Applied data mining for business decision making using R*. CRC Press.

Raju, D. and Schumacker, R. (2015) ‘Exploring Student Characteristics of Retention That Lead To Graduation in’, *Journal of College Student Retention: Research, Theory & Practice*, 16(4), pp. 563–591. doi: 10.2190/CS.16.4.e.

‘RapidMiner Studio 9.2’ (2019).

Rashu, R. I., Haq, N. and Rahman, R. M. (2003) ‘Data mining approaches to predict final grade by overcoming class imbalance problem’, *2014 17th International Conference on Computer and Information Technology, ICCIT 2014*, (March), pp. 14–19. doi: 10.1109/ICCITechn.2014.7073095.

Romero, C. and Ventura, S. (2010) ‘Educational data mining: a review of the state of the art’, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. Ieee, 40(6), pp. 601–618.

Rüping, S. (2000) ‘MySVM’.

Rüping, S. (2003) ‘myKLR-kernel logistic regression. University of Dortmund, Department of Computer Science’.

Shahiri, A. M., Husain, W. and Rashid, N. A. (2015) ‘A Review on Predicting Student’s Performance Using Data Mining Techniques’, *Procedia Computer Science*. Elsevier Masson SAS, 72, pp. 414–422. doi: 10.1016/j.procs.2015.12.157.

Shlens, J. (2014) ‘A Tutorial on Principal Component Analysis’. doi: 10.1.1.115.3503.

Stapel, M., Zheng, Z. and Pinkwart, N. (2016) ‘An Ensemble Method to Predict Student Performance in an Online Math Learning Environment’, *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 231–238.

Thammasiri, D. *et al.* (2014) ‘A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition’, *Expert Systems with Applications*. Elsevier Ltd, 41(2), pp. 321–330. doi: 10.1016/j.eswa.2013.07.046.

Tharwat, A. (2018) ‘Classification assessment methods’, *Applied Computing and Informatics*. The Author. doi: 10.1016/j.aci.2018.08.003.

Tinto, V. (2007) ‘Taking student retention seriously’, *Annual Recruitment and Retention Conference*, pp. 1–8. Available at: http://survey.csuprojects.org/uploads/a-/nu/a-nuQmE5d6vFwnkDnNNn7Q/Tinto-re-Taking-Student-Retention-Seriously.pdf.

Viera, A. J. and Garrett, J. M. (2005) ‘Vierra_2005 _Interrater_agreement.Kappa_statistic’, (May), pp. 360–363. doi: Vol. 37, No. 5.

Yang, S., Guo, J. Z. and Jin, J. W. (2018) ‘An improved Id3 algorithm for medical data classification’, *Computers and Electrical Engineering*. Elsevier Ltd, 65, pp. 474–487. doi: 10.1016/j.compeleceng.2017.08.005.

Zou, Q. *et al.* (2016) ‘Finding the Best Classification Threshold in Imbalanced Classification’, *Big Data Research*. Elsevier Inc., 5, pp. 2–8. doi: 10.1016/j.bdr.2015.12.001.

Appendix 1

Literature Review Areas

Summary of literature review papers

	Year	Title	Class	DT	k-NN	ID3	NB	NN	LR	SVM	GBT	RF	En	Bag	Boost	Vote 2	Vote 3	Imbalance	RUS	ROS	Weighting	SMOTE
1	**2018**	2018 ASEE Southeastern Section Conference Predicting Academic Achievement in Fundamentals of Thermodynamics using Supervised Machine Learning Techniques	1	1								1										
2	**2018**	An efficient use of ensemble methods to predict students academic performance	1		1	1	1			1			1									
3	**2018**	Analyze and Predict Student Dropout from Online Programs	1	1	1		1		1	1		1										
4	**2018**	Classification based data mining algorithms to predict slow, average and fast learners in educational system using Weka.	1				1		1			1										
5	**2018**	Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout	1						1													
6	**2018**	Predicting and Analysis of Students' Academic Performance using Data Mining Techniques	1									1										
7	**2018**	Predicting student academic performance using multi-model heterogeneous ensemble approach	1	1				1		1			1									
8	**2017**	Analyzing undergraduate students' performance using educational data mining	1	1																		
9	**2017**	Classification and prediction based data mining algorithms to predict students' introductory programming performance	1	1			1	1	1	1												
10	**2017**	Early Prediction of Student Success: Mining Students Enrolment Data	1	1																		
11	**2017**	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	1	1			1	1		1			1									
12	**2017**	Predicting student performance with Neural Networks	1	1	1		1	1	1		1											
13	**2016**	An Ensemble Method to Predict Student Performance in an Online Math Learning Environment	1	1			1		1			1	1			1						
14	**2016**	Educational Data Mining & Students' Performance Prediction	1	1			1												1			
15	**2016**	Predicting Student Dropout in Higher Education	1		1				1			1										
16	**2016**	Prediction of students performance using Educational Data Mining	1				1															
17	**2016**	Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students	1	1	1		1		1	1												
18	**2015**	A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes	1	1					1	1		1			1							
19	**2015**	A multivariate approach to predicting student outcomes in web-enabled blended learning courses	1						1													
20	**2015**	A Review on Predicting Student's Performance Using Data Mining Techniques	1	1	1		1	1		1												
21	**2015**	Analyzin early At Risk factors in higher edcuation e-learning	1	1			1		1													
22	**2015**	Classification and prediction based data mining algorithms to predict slow learners in education sector	1	1			1															
23	**2015**	Data mining approaches to predict final grade by overcoming class imbalance problem	1	1			1	1										1	1	1		1
24	**2015**	Exploring Student Characteristics of Retention That Lead To Graduation in	1	1				1	1													
25	**2015**	OU Analyse : Analysing at - risk students at The Open University	1	1			1															
26	**2014**	A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition	1	1				1	1	1								1	1	1		1
27	**2014**	Building a Student At-Risk Model : An End-to-End Perspective	1	1			1	1	1				1									
28	**2014**	Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative	1	1			1		1	1									1			
29	**2013**	Predicting Student Performance by Using Data Mining Methods for Classification	1	1	1		1															
30	**2010**	A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education	1				1						1									
		Total number of papers	30	22	7	1	19	9	15	10	1	7	6	0	1	1	0	2	4	2	0	2
		% of papers in each category	100	73.33	23.33	3.33	63.33	30.00	50.00	33.33	3.33	23.33	20.00	0.00	3.33	3.33	0.00	6.67	13.33	6.67	0.00	6.67

Appendix 2

Model Results – Sampling and Weighting

Original Model

	LR	GBT	SVM	NN	NB	RF	DT
AUC	0.953	0.937	0.913	0.798	0.741	0.689	0.662
Accuracy	0.908	0.870	0.810	0.742	0.700	0.692	0.694
G-mean	0.905	0.871	0.819	0.734	0.69	0.592	0.590
Kappa	0.814	0.739	0.635	0.478	0.388	0.338	0.341
f-measure	0.895	0.858	0.818	0.697	0.649	0.516	0.512
Specificity	92.99	86.22	72.43	80.61	75.82	94.87	95.77
Sensitivity	88.12	88.07	92.67	66.77	62.71	37	36.29
Precision	90.96	83.07	73.49	74.12	67.41	85.28	87.33
False Positive FP	9.93	19.53	117.2	41.2	51.4	7.27	9
False Negative FN	13.4	13.47	24.8	56.2	63.1	71.07	107.8
Classify error	0.092	0.130	0.19	0.258	0.300	0.308	0.306

SMOTE Sampling

	GBT	LR	SVM	RF	NN	DT	NB
AUC	0.958	0.947	0.942	0.775	0.758	0.740	0.740
Accuracy	0.910	0.891	0.910	0.711	0.712	0.718	0.695
G-mean	0.910	0.889	0.909	0.632	0.697	0.730	0.685
Kappa	0.819	0.779	0.818	0.391	0.410	0.436	0.378
f-measure	0.899	0.877	0.899	0.566	0.653	0.721	0.645
Specificity	91.72	90.45	92.24	95.06	78.69	76.30	74.83
Sensitivity	90.19	87.41	89.54	42.07	61.81	69.85	62.71
Precision	89.68	88.03	90.31	87	70.81	74.63	66.52
False Positive FP	35.20	13.53	11.00	10.60	45.30	33.60	53.50
False Negative FN	33.2	14.2	11.8	98	64.6	42.73	63.1
Classify error	0.090	0.109	0.089	0.289	0.288	0.212	0.305

Bootstrap Sampling

	GBT	LR	SVM	RF	NN	NB	DT
AUC	0.932	0.863	0.819	0.796	0.771	0.747	0.661
Accuracy	0.863	0.790	0.784	0.708	0.723	0.696	0.694
G-mean	0.857	0.778	0.781	0.627	0.721	0.687	0.591
kappa	0.721	0.569	0.563	0.376	0.444	0.381	0.342
f-measure	0.840	0.748	0.767	0.519	0.689	0.648	0.513
Specificity	90.03	85.65	80.62	93.74	75.54	74.36	95.67
Sensitivity	81.67	70.58	75.59	41.91	68.86	63.54	36.46
Precision	86.81	79.68	76.15	84.26	69.5	66.45	87.11
False Positive FP	14.33	30.50	41.20	8.87	52.00	54.50	9.20

False Negative FN	20.66	182.10	41.30	65.53	52.70	61.70	107.50
Classify error	0.137	0.210	0.216	0.292	0.274	0.304	0.306

Stratified Sampling

	GBT	SVM	LR	NN	RF	NB	DT
AUC	0.922	0.883	0.821	0.817	0.780	0.741	0.658
Accuracy	0.844	0.834	0.752	0.764	0.716	0.702	0.691
G-mean	0.830	0.835	0.727	0.750	0.632	0.691	0.589
Kappa	0.679	0.671	0.486	0.515	0.395	0.391	0.334
f-measure	0.810	0.816	0.685	0.715	0.568	0.650	0.510
Specificity	91.67	85.89	86.21	83.81	95.01	76.25	95.01
Sensitivity	75.230	81.09	61.23	67.07	42.07	62.53	36.46
Precision	87.810	82.28	78.09	77.04	87.87	67.73	85.61
False Positive FP	35.40	20	29.10	22.93	9.90	50.50	10.60
False Negative FN	83.800	21.33	65.60	37.13	98	63.40	107.50
Classify error	0.156	0.166	0.248	0.236	0.284	0.298	0.309

Weighting Stratified

	GBT	SVM	NN	NB	DT
AUC	0.966	0.835	0.796	0.741	0.660
Accuracy	0.931	0.712	0.737	0.690	0.693
G-mean	0.930	0.717	0.733	0.704	0.589
kappa	0.861	0.431	0.467	0.372	0.338
f-measure	0.922	0.684	0.703	0.651	0.511
Specificity	94.31	65.31	76.28	75.97	95.49
Sensitivity	91.66	78.63	70.44	65.26	36.29
Precision	92.79	70.93	70.83	64.99	86.71
False Positive FP	8.07	73.70	50.40	39.73	9.60
False Negative FN	9.40	36.20	50	39.20	107.80
Classify error	0.069	0.288	0.263	0.310	0.307

Appendix 3

Early Prediction Model results

Course data Quiz 1 all courses							
	AUC	**GM**	**f-measure**	**Accuracy**	**kappa**	**Precision**	**Recall**
GBT with Weighting	0.868	0.803	0.832	0.802	0.592	87.59	79.63
GBT with SMOTE	0.864	0.792	0.836	0.8	0.579	85.37	82.22
Logistic Regression	0.857	0.800	0.85	0.814	0.604	85.36	85.19
SVM with Bagging	0.828	0.764	0.823	0.782	0.535	82.76	82.96
LR SVM NN	0.774	0.775	0.832	0.793	0.561	83.02	84.07
SVM with Boost	0.695	0.721	0.799	0.745	0.456	79.11	81.11
Course data Quizzes all courses							
	AUC	**GM**	**f-measure**	**Accuracy**	**kappa**	**Precision**	**Recall**
GBT with Weighting	0.893	0.807	0.825	0.8	0.594	89.28	77.41
GBT with SMOTE	0.892	0.816	0.839	0.812	0.613	88.9	79.63
Logistic Regression	0.875	0.781	0.846	0.805	0.578	82.96	86.67
SVM with Bagging	0.839	0.775	0.838	0.795	0.56	83.14	84.81
LR SVM NN	0.76	0.766	0.83	0.786	0.541	82.43	82.07
SVM with Boost	0.5	0.744	0.817	0.77	0.505	80.53	83.33
Course data Quizzes and Practical assessment all courses							
	AUC	**GM**	**f-measure**	**Accuracy**	**kappa**	**Precision**	**Recall**
GBT with Weighting	0.903	0.808	0.852	0.818	0.616	86.32	84.81
GBT with SMOTE	0.892	0.804	0.832	0.802	0.593	87.53	79.63

Logistic Regression	0.881	0.766	0.848	0.802	0.566	81.35	88.89
SVM with Bagging	0.871	0.787	0.854	0.814	0.596	83.21	88.15
LR SVM NN	0.776	0.795	0.859	0.821	0.612	83.68	88.52
SVM with Boost	0.5	0.753	0.837	0.788	0.534	80.97	87.04
Course data Quizzes, Practical and Project assessments all courses							
	AUC	**GM**	**f-measure**	**Accuracy**	**kappa**	**Precision**	**Recall**
GBT with Weighting	0.911	0.832	0.863	0.835	0.653	89.24	84.07
GBT with SMOTE	0.885	0.814	0.841	0.812	0.609	88.85	80.37
Logistic Regression	0.877	0.740	0.85	0.798	0.544	79.04	92.22
SVM with Bagging	0.871	0.758	0.853	0.805	0.565	80.32	91.11
LR SVM NN	0.766	0.778	0.862	0.818	0.598	81.74	91.48
SVM with Boost	0.5	0.749	0.842	0.793	0.541	80.06	89.26
Course data Coursework and Final assessments all courses							
	AUC	**GM**	**f-measure**	**Accuracy**	**kappa**	**Precision**	**Recall**
GBT with Weighting	0.923	0.840	0.880	0.851	0.682	88.30	88.15
Logistic Regression	0.913	0.746	0.869	0.818	0.584	79.00	96.67
GBT with SMOTE	0.910	0.836	0.855	0.830	0.651	90.54	81.11
SVM with Bagging	0.909	0.760	0.873	0.825	0.602	79.92	96.30
LR SVM NN	0.781	0.769	0.878	0.832	0.617	80.64	96.67
SVM with Boost	0.500	0.75725	0.873	0.825	0.6	79.64	96.67

Appendix 4

Decision Tree Description

```
High School Average > 89.150
|   CEPA > 152.190
|   |   High School Average > 97.900: At Risk {At Risk=2, Good Standing=0}
|   |   High School Average ≤ 97.900
|   |   |   High School MATH > 97.770
|   |   |   |   High School ENGL > 92.802: Good Standing {At Risk=0, Good Standing=6}
|   |   |   |   High School ENGL ≤ 92.802
|   |   |   |   |   High School ENGL > 90.028: At Risk {At Risk=13, Good Standing=0}
|   |   |   |   |   High School ENGL ≤ 90.028: Good Standing {At Risk=1, Good Standing=5}
|   |   |   High School MATH ≤ 97.770: Good Standing {At Risk=84, Good Standing=583}
|   CEPA ≤ 152.190: At Risk {At Risk=12, Good Standing=1}
High School Average ≤ 89.150
|   CEPA > 209: Good Standing {At Risk=0, Good Standing=25}
|   CEPA ≤ 209: At Risk {At Risk=2014, Good Standing=1072}
```

Gradient Boosted Trees Description

```
High School Average > 89.250
|   CEPA > 152.299
|   |   High School Average > 97.350: At Risk {At Risk=2, Good Standing=0}
|   |   High School Average ≤ 97.350
|   |   |   High School ENGL > 73.607
|   |   |   |   High School MATH > 97.627
|   |   |   |   |   High School ENGL > 91.994: Good Standing {At Risk=0, Good Standing=7}
|   |   |   |   |   High School ENGL ≤ 91.994
|   |   |   |   |   |   Gender Code > 1.500
|   |   |   |   |   |   |   Campus Code > 7: Good Standing {At Risk=0, Good Standing=2}
|   |   |   |   |   |   |   Campus Code ≤ 7: At Risk {At Risk=2, Good Standing=0}
|   |   |   |   |   |   Gender Code ≤ 1.500: At Risk {At Risk=11, Good Standing=0}
|   |   |   |   High School MATH ≤ 97.627: Good Standing {At Risk=41, Good Standing=382}
|   |   |   High School ENGL ≤ 73.607: At Risk {At Risk=2, Good Standing=0}
|   CEPA ≤ 152.299: At Risk {At Risk=6, Good Standing=0}
High School Average ≤ 89.250
|   CEPA > 209: Good Standing {At Risk=0, Good Standing=17}
|   CEPA ≤ 209
|   |   High School MATH > 95.304: Good Standing {At Risk=0, Good Standing=10}
|   |   High School MATH ≤ 95.304
|   |   |   High School ENGL > 92.500: Good Standing {At Risk=0, Good Standing=4}
|   |   |   High School ENGL ≤ 92.500: At Risk {At Risk=1352, Good Standing=749}
```

Random Forest Tree Description

```
High School Average > 89.250
|   CEPA > 152.190
|   |   High School ENGL > 72.600
|   |   |   Absence % > 9.725
|   |   |   |   CEPA > 184.500: Good Standing {At Risk=0, Good Standing=6}
|   |   |   |   CEPA ≤ 184.500
|   |   |   |   |   High School Average > 89.950
|   |   |   |   |   |   High School Average > 90.200
|   |   |   |   |   |   |   IELTS Band > 5.250: At Risk {At Risk=2, Good Standing=0}
|   |   |   |   |   |   |   IELTS Band ≤ 5.250
|   |   |   |   |   |   |   |   High School ENGL > 88.862: At Risk {At Risk=2, Good Standing=2}
|   |   |   |   |   |   |   |   High School ENGL ≤ 88.862: Good Standing {At Risk=0, Good Standing=14}
|   |   |   |   |   |   High School Average ≤ 90.200: At Risk {At Risk=10, Good Standing=0}
|   |   |   |   |   High School Average ≤ 89.950: Good Standing {At Risk=0, Good Standing=12}
|   |   |   Absence % ≤ 9.725
|   |   |   |   High School Average > 97.350: At Risk {At Risk=1, Good Standing=0}
|   |   |   |   High School Average ≤ 97.350
|   |   |   |   |   High School Average > 96.250
|   |   |   |   |   |   High School Average > 96.550: Good Standing {At Risk=0, Good Standing=2}
|   |   |   |   |   |   High School Average ≤ 96.550: At Risk {At Risk=1, Good Standing=0}
|   |   |   |   |   High School Average ≤ 96.250
|   |   |   |   |   |   High School Average > 90.850
|   |   |   |   |   |   |   High School Average > 90.950
|   |   |   |   |   |   |   |   IELTS Band > 4.645: Good Standing {At Risk=7, Good Standing=167}
|   |   |   |   |   |   |   |   IELTS Band ≤ 4.645: Good Standing {At Risk=5, Good Standing=7}
|   |   |   |   |   |   |   High School Average ≤ 90.950
|   |   |   |   |   |   |   |   High School ENGL > 91.484: Good Standing {At Risk=0, Good Standing=3}
|   |   |   |   |   |   |   |   High School ENGL ≤ 91.484: At Risk {At Risk=3, Good Standing=0}
|   |   |   |   |   |   High School Average ≤ 90.850: Good Standing {At Risk=0, Good Standing=98}
|   |   High School ENGL ≤ 72.600: At Risk {At Risk=1, Good Standing=0}
|   CEPA ≤ 152.190: At Risk {At Risk=7, Good Standing=0}
High School Average ≤ 89.250
|   High School MATH > 95.347: Good Standing {At Risk=0, Good Standing=7}
|   High School MATH ≤ 95.347
|   |   CEPA > 209: Good Standing {At Risk=0, Good Standing=11}
|   |   CEPA ≤ 209
|   |   |   IELTS Band > 6.081
|   |   |   |   Absence % > 6
|   |   |   |   |   CEPA > 200.500
|   |   |   |   |   |   CEPA > 206: At Risk {At Risk=3, Good Standing=0}
|   |   |   |   |   |   CEPA ≤ 206: Good Standing {At Risk=0, Good Standing=2}
|   |   |   |   |   CEPA ≤ 200.500
|   |   |   |   |   |   IELTS Band > 6.098: At Risk {At Risk=10, Good Standing=0}
|   |   |   |   |   |   IELTS Band ≤ 6.098
|   |   |   |   |   |   |   CEPA > 195.500: At Risk {At Risk=2, Good Standing=0}
|   |   |   |   |   |   |   CEPA ≤ 195.500: Good Standing {At Risk=0, Good Standing=1}
|   |   |   |   Absence % ≤ 6
|   |   |   |   |   Absence % > 2.868: Good Standing {At Risk=0, Good Standing=14}
|   |   |   |   |   Absence % ≤ 2.868
|   |   |   |   |   |   High School Average > 74
|   |   |   |   |   |   |   CEPA > 169.830: Good Standing {At Risk=0, Good Standing=8}
|   |   |   |   |   |   |   CEPA ≤ 169.830: At Risk {At Risk=1, Good Standing=0}
|   |   |   |   |   |   High School Average ≤ 74: At Risk {At Risk=2, Good Standing=0}
|   |   |   IELTS Band ≤ 6.081
|   |   |   |   Absence % > 14.665: At Risk {At Risk=41, Good Standing=0}
|   |   |   |   Absence % ≤ 14.665
|   |   |   |   |   High School Average > 60.900
|   |   |   |   |   |   High School ENGL > 58.250
|   |   |   |   |   |   |   High School MATH > 59.350
|   |   |   |   |   |   |   |   CEPA > 191.500: Good Standing {At Risk=11, Good Standing=27}
|   |   |   |   |   |   |   |   CEPA ≤ 191.500: At Risk {At Risk=1107, Good Standing=517}
|   |   |   |   |   |   |   High School MATH ≤ 59.350: At Risk {At Risk=7, Good Standing=0}
|   |   |   |   |   |   High School ENGL ≤ 58.250: Good Standing {At Risk=0, Good Standing=1}
|   |   |   |   |   High School Average ≤ 60.900: Good Standing {At Risk=0, Good Standing=4}
```