# Introduction

In this article “APPLYING PREDICTIVE ANALYTICS IN IDENTIFYING STUDENTS AT RISK: A CASE STUDY”, a case study was attempting to find an institutional modeling project using an appropriate learning algorithm to predict whether students dropping out between their first and second year of study could be intervened with using strategies for intervention to assist the students at risk and improve success rates of completing their first and second years of college.

Institutions were attempting to find the explanation for students who are considered, “at risk” for dropping out of the educational institutions early and to prevent a loss of retention for these “at risk” students. This problem was extremely important for the institution, as it helped to reduce the number of dropouts between first and second year. This problem being solved also induced other programs, such as the “First-Year Experience,” securing funding from the Teaching Development Grant from the Department of Higher Education and Training (DHET), to assist in the reduction of dropouts between first and second-year college students as stated previously.

Data was collected from the Department of Institutional Planning beginning in 2014, shortly after the appointment of the vice-chancellor using the project called “Know Our Students”. The data acquired was collected with the understanding that the project was to collect as much information about the characteristics of the students within the institution. Reports regarding the demographic profiles of the students including their high school and its quintile, along with other identifiable characteristics of the students’ behaviors, mindset, and other such personal identifiers, as to accurately collect the information to determine if the data collection was appropriate and accurate for the project.

# Methods and Results

The team responsible for organizing and collecting the data, along with interpreting the results from the algorithms, apply the Cross-Industry Standard Process Model for Data Mining also known as KRISP-DM. These six steps included the following exercises:

* Business understanding
* Data understanding
* Data preparation
* Modeling
* Evaluation
* Deployment

A total of 1,593 student records were used after all missing data had been either removed or binned. In all, 28% of these records indicated that the student was classified as second-year dropout. This equates to 452 students identified. This data was imported from Higher Education Data Analyzer (HEDA) management system. This system allowed for the use of an automated tool for predictive analysis, using the capabilities of Structured Query Language, which was already apart of the HEDA infrastructure. To obtain better understanding of the metrics and data collection, some points were raised after a workshop held with other teams involved in the project, the following became clear:

* There is no single algorithm that provided reliable predictions across a range of qualifications.
* Cape Peninsula University of Technology in Cape Town would need to continue to gather additional information about incoming students in order to apply predictive analytics before the students began at the institution.
* Tools already used by the institution could help provide real time descriptive analytics based on student engagement with the e-learning system, in which case the institution was using Blackboard.
* This was not without challenge, as it would be difficult to produce predictive analytics on a particular types of data points that were required to make an accurate prediction based on the analytics that came in through the e-learning platform.

From this workshop it was decided to focus on the national diploma in it for the case study. Institutional operational data for cohorts of first time entering students enrolled in the qualification at Cape peninsula University of Technology from 2008 to 2014 were used in the analysts to predict student dropout by the second year of study. KNIME was used to perform the predictive modeling, as the dependent slash target variable the second-year dropout students were classified in the following way:

* If a student did not return in the second year of study, he or she was coded as one
* If a student did not return in the second year of study, he or she was coded as a zero.

A plethora of information on background and demographics, and performance linked data from first semester data from Cape peninsula University of Technology was extracted, prepared and cleaned in HEDA and used to create the model.

A total of 27 variables were used to test collinearity, after which 22 variables were included in the descriptive analysis. From this process only selected variables were used in the analysis, since redundant variables were excluded, and some variables had to be transformed or combined to prevent overfitting of the model. The logistic regression, Naïve Bayes, and decision tree algorithms were used for this study. The linear regression technique was used to model the relationship between a binary dependent variable and categorical and slash or continuous independent/predictor variable. Naïve Bayes Classifiers were used to handle an arbitrary number of independent variables whether continuous or categorical in nature. The decision tree was used to build classification or regression models in the form of a tree structure. It was used to handle missing data automatically and visually and intuitively support the results from the model. All the algorithms were used against randomly subdivided training and test data sets. The data set was used with information collected from 2008 to 2013. The 2014 data set was kept aside for later use as a validation data set. The data sets were randomly subdivided into 70% training and 30% testing data sets. Data sets for training accounted for 988 records of first time entered students which was being used to build the models in order to identify the predictor variables. The remaining set of records, or 424 records were used as the testing data set to assume the accuracy of the models. The 2014 data set used 181 records to predict the outcome based on the selected model. the logistic regression, Naïve Bayes, and decision tree nodes were applied to eight independent variables:

* Technical programming
* Development software
* Information systems
* Information technical skills
* System software
* Financial aid
* Grade 12 mathematics
* Type of accommodation

After conducting the analyses, the logistic regression score retained four significant predictor variables, indicating that students enrolled with the national diploma in it who did not receive any bursary and had low marks in the data science, information systems and information technology system modules were more likely to drop out by the second year of study. From the confusion matrix it was clear the logistic regression model resulted to an overall percentage of correct classification of 88% based on the accounts in each quadrant of each predicted probability threshold. The actual second year dropout rate for students classified correctly at 73.3%. The model performed well on the logistic regression model as it had a high percentage accuracy of 88.6% with only an 11.3% error rate and was subsequently used on the validation testing set.

# Conclusion

This study was important for the university to identify potential at risk students early in the first year of study in order to implement support and intervention programs to improve the retention rate for a student continuing into the second year. Using the predictive analytics case studies, it was demonstrated that through the algorithms on first year students enrolled and the national diploma information technology at Cape Peninsula University of Technology, it helped to identify and help alleviate the potential of students that were considered at risk from becoming a statistic of the recurring revolving door of higher education, while helping to guide students towards graduation, but also provided a benefit to the students family, to the community, and to the economy. These models also had an economic incentive for the university, as it helps to identify those students who are potential at risk of not continuing to their second year.