Comparing the Viability of Different Machine Learning Models to Predict Student Retention

Tribikram Adhikari

Department of Computer Science,

Graceland University

1 University Pl, Lamoni, IA, 50140

tribikram@sting.graceland.edu

Abstract

This study compares the effectiveness of three machine learning models: logistic regression, naïve Bayes and decision trees in predicting student retention. Student retention is a key gauge of institutional success for colleges. It brings more tuition revenue without costly expansion of college recruiting as well as increases its reputation through high graduation rates. This study uses student data collected during their enrollment at Graceland University to predict whether they will successfully graduate from the university or not. Feature engineering and feature selection methods were carried out to identify and extract the most useful features from the data set. The data set was then split into training and the holdout set to train and test the effectiveness of the machine learning models. Evaluation of the models was done through well-known indicators like precision, recall, and f-measure. In this study, logistic regression proved to be the most effective way to predict student retention. Finally, this study concludes by highlighting its shortcomings and discussing potential ways of future improvement.

Keywords: Data science, student retention, decision tree, logistic regression, naïve Bayes

1. Introduction

Student retention is a matter of concern for academic institutions as it is a key indicator of their performance. An increase in retention would bring more tuition revenue without costly expansion of college recruiting. It would also lead to a higher graduation rate, strengthening the school's reputation. Therefore, early identification of students who are prone to dropping out is important for the success of any retention strategy.

Studies in the past have shown that precollege academic achievements of students like high school GPA, ACT/SAT scores, can be useful predictors of student retention (Maurtaugh, Burns, Schuster pg. 356), however, other characteristics like students' mental and emotional health, family education history, etc. have also been found to have an impact on students' motivation to complete college (Ishitani pg. 434). Clearly, it would be beneficial for institutions to have a mechanism to predict student retention considering these factors.

This research compares the outcomes of three machine learning models: Logistic Regression, Naïve Bayes Classifier, and Decision Trees in predicting student retention. The data used in this study was received from the admissions department at Graceland University which had been collected from over 1700 students during the span of 2007 to 2014. During the training period, the data was first examined for outliers and missing values. The missing values were then filled with appropriate values that would reduce the scale of incorrect predictions. Relevant features for training the models were selected through feature selection and feature extraction process. Finally, the models were compared using well-known indicators like Precision, Recall, and F-measure.

2. Literature Review

Research related to student retention at academic institutions can be traced back over 70 years (Reason pg. 172). Early studies have found that more than 40% of all college entrants leave without earning a degree, 75% of which drop out in the first 2 years of school (Gerdes and Mallinckrodt pg. 281). Until the 1970s, the reason behind low student retention was believed to be a failure on the part of the student and not the institution (Tinto pg. 2). However, the belief that the sole reason for students' dropping out of higher education is due to their inability to cope with college stress and lack of willingness to succeed was debunked by research carried out during the 1980s (Tinto pg. 3). These studies focused on factors like teacher-student relationship, student participation in extracurricular activities, and others that helped students get accustomed to college life (Tinto pg. 3). The transition to college is often marked by emotional, social and academic challenges (Gerdes and Mallinckrodt pg. 281). Studies that focused on students' academic ability and retention found that academic ability explained no more than half of the variance in their decision to drop out (Gerdes and Mallinckrodt pg. 281).

In the past, there have been several attempts to model student retention. A conceptual model known as the Input-Environment-Output (I-E-O) model was created in 1991. According to this model, input characteristics and educational environment should be weighed more than the outcomes

(Alkasawneh and Hargraves pg. 35). Input characteristics refer to individual student's race/ethnicity, gender, family's educational background, college admission test scores, high school GPA and other self-reported data like goals and college expectations (Alkasawneh and Hargraves pg. 35). These input characteristics of a student have a strong influence on the educational environment. The educational environment referred to everything a student experienced academically and socially during college. This model showed that lack of involvement in college was a major cause of student dropout.

A study conducted at Oregon State University used survival analysis to model student retention based on the same input characteristics. The study showed a strong positive relationship between student retention and features like high school GPA, residency, first-quarter college performance, ethnicity/race and enrollment in their Freshman Orientation Course (Maurtaugh, Burns and Schuster pg. 369). The study also found an indirect relationship between age at enrollment and retention; as there was an increase in age, there was a decrease in retention (Maurtaugh, Burns and Schuster pg. 369).

This research uses similar approaches to predict student retention through some well-known statistical models. The data used in this research was provided by the admissions department at Graceland University and contained several features that were determined as good predictors of student attrition by previous studies discussed in this literature review.

2.1 Predictive Models

This section discusses the different models that were used in this study to predict student retention based on the data provided by Graceland University. The models that were selected for comparison were Logistic Regression, Naïve Bayes and Decision Trees.

2.1.1 Logistic Regression

In machine learning, logistic regression is used to find the best fitting model to describe the relationship between the categorical characteristics of the dependent variable and a set of independent variables (Yan and Lee pg. 913). The output of logistic regression is the probability of an event, constrained between 0 and 1.

Let x be an event and $P(x)$ be the probability of that event occurring then the logistic function is:

$$P(x) = \frac{1}{1 + e^{-g(x)}} = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

The logistic or logit model is:

$$Logit = g(x) = \ Log\frac{P(x)}{1-P(x)} = \ \alpha + \ \beta_1 x_1 \ + \ \beta_2 x_2 + \cdots + \ \beta_k x_k$$

In order to compute $P(x)$, the parameters α and $\beta_1 \ldots . \beta_k$ need to be determined in advance. These parameters are computed using the maximum likelihood method (Houston and Woodruff pg. 3).

2.1.2 Naïve Bayes Classifier

Naive Bayes classifier depends on the Bayes' theorem:

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

where A and B are two events and $P(A)$ and $P(B)$ are the probabilities of these two events occurring respectively. $P(A\,|B)$ represents the probability of event A occurring given that event B has already occurred (Islam, Wu and Ahmadi pg. 134). A property of Naïve Bayes classification is that it assumes independence between the features. This means that the occurrence of one event does not affect the probability of the other.

In the example below, Bayes' theorem is applied to classify a student having a 3.0 GPA and ACT 18 using the data given in Table 2.

GPA	ACT	Outcome (0 – dropped out, 1 – successfully retained)
2.0	18.0	1
3.0	20.0	1
2.0	22.0	0
4.0	18.0	0
3.0	22.0	0

Table 1: Example data for Naïve Bayes classification

Let O be the outcome then we have:

$$P\ (O = 1|GPA = 3, ACT = 18) = \frac{P(GPA = 3, ACT = 18\ |O = 1)\ P(O = 1)}{P(\ GPA = 3, ACT = 18)}$$

By applying the chain rule of conditional probability to the numerator, we get:

$$P\ (O = 1|GPA = 3, ACT = 18) = \frac{P(GPA = 3|O = 1)\ P(ACT = 18\ |O = 1)\ P(O = 1)}{P(\ GPA = 3, ACT = 18)}$$

Calculating the numerator, we get:

$$P\ (O = 1|GPA = 3, ACT = 18) = \frac{0.5\ \cdot 0.5\ \cdot\ 0.4}{P(\ GPA = 3, ACT = 18)}$$

$$P\ (O = 1|GPA = 3, ACT = 18) = \frac{0.1}{P(\ GPA = 3, ACT = 18)}$$

Similarly, we carry out the same operations for $P\ (O = 0|GPA = 3, ACT = 18)$ we get:

$$P\ (O = 0|GPA = 3, ACT = 18) = \frac{0.33\ \cdot 0.33\ \cdot\ 0.6}{P(\ GPA = 3, ACT = 18)}$$

$$P\ (O = 0|GPA = 3, ACT = 18) = \frac{0.067}{P(\ GPA = 3, ACT = 18)}$$

Now both the probabilities are compared to each other. Since the final expressions for both the probabilities i.e. $P\ (O = 1|GPA = 3, ACT = 18)$ and $P\ (O = 0|GPA = 3, ACT = 18)$, have the same denominator $P(\ GPA = 3, ACT = 18)$, thus, by comparison, the probability of being retained is greater than that of dropping out. In this example, only some combinations had the probabilities. In an actual setting, a distribution is required.

2.1.3 Decision Trees

Decision trees are graph structures that use a tree-like model of decisions and the possible outcomes (Quinlan pg. 3). Figure 1 shows an example of a decision tree.

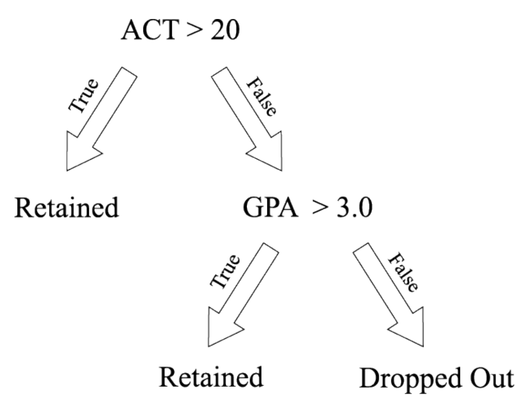

Figure 1: Example of a decision tree

The decision tree in the figure above predicts whether a student is going to be retained or not by looking at ACT score and GPA. The labels “True” and “False” in the edges indicate whether the “ACT > 20” and “GPA > 3.0” conditions are met.

Decision trees use recursive partitioning to partition data in appropriate values until a tree structure has emerged (Strobl, Tutz and Malley pg. 330). The decision tree algorithm tries to find a way to partition the data such that the parts are as homogeneous as possible (Strobl, Tutz and Malley pg. 330). In case a fully homogeneous part is not possible, the most common value is selected. The decision tree based on the data below demonstrates this process.

ACT	GPA	Outcome (0 – dropped out, 1 – successfully retained)
15	3.0	0
15	3.6	1
22	2.0	0
22	3.0	1
28	4.0	1

Table 2: Example decision tree data

Like the previous example, the goal is to predict whether a student is going to be retained or not using ACT score and GPA. As the data contains only two independent variables, it can be visualized as a scatterplot.

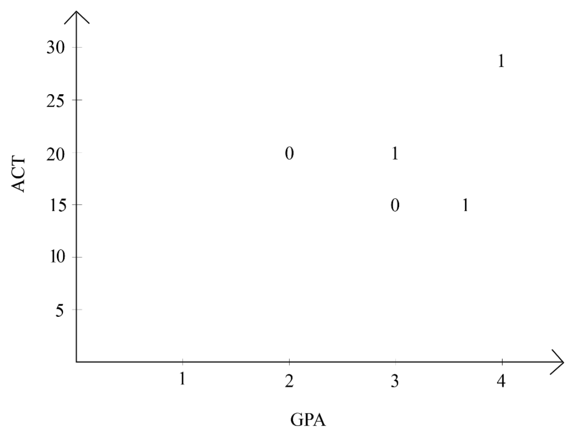

Figure 2: Scatter plot of student data

The axes X and Y represent the independent variables, whereas the points (Retained = 1 and Dropped out = 0) represent the dependent variables. The plot can be portioned using a decision tree algorithm. This process is demonstrated in Figure 3.

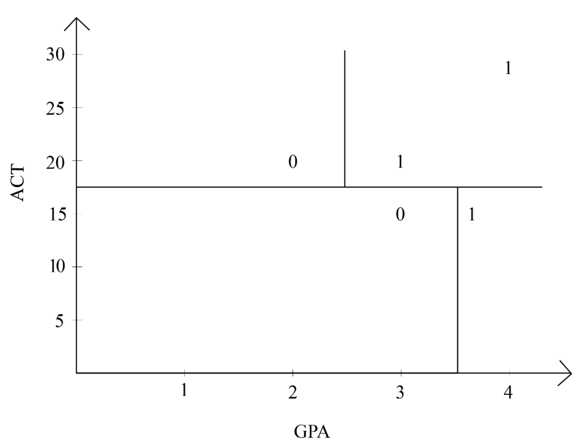

Figure 3: Partitioned scatter plot.

The partition locations and the number of partitions are determined by the decision tree algorithm. The partitioned plot can be represented in the form of a decision tree as shown in Figure 6.

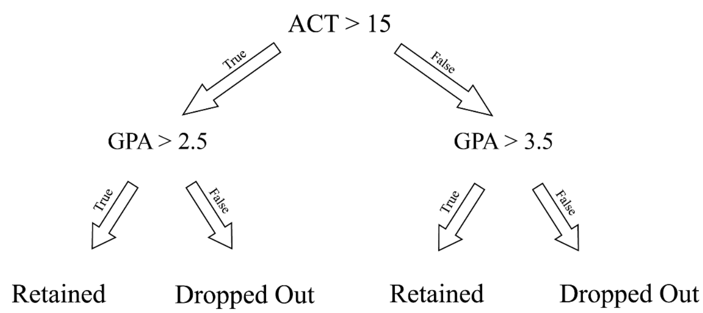

Figure 4: scatter plot represented in the form of a decision tree.

2.2 Evaluation Methods

Evaluation of the models is done by comparing the predicted values with the actual values. Often, a confusion matrix is used to describe the performance of the classifier (Fawcett pg. 862). An example of a confusion matrix is given below:

	Predicted as True	Predicted as False
Actually True	True Positives	False Negative
Actually False	False Positives	True Negative

Table 3: Confusion matrix with possible prediction results

In the table above, true positives (TP) refer to the cases which were predicted to be positive (in this context a student going to be successfully retained), and they were retained, true negatives (TN) refers to the cases which were predicted to be false (in this context a student going to drop out) and they did drop out, false positives (FP) refers to the cases where the predictions were positive but in reality the results were negative, finally, false negatives (FN) refers to the cases where the predictions were negative but the real results were positive (Fawcett pg. 862).

Accuracy, the ratio of correct predictions to total predications, can be calculated from a confusion matrix. It is represented as:

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

However, accuracy does not show how the minority class is classified (Fawcett pg. 862), for example, consider a data set of 100 students, of which 90 were successfully retained, if the model predicts that all students were retained it will still have an accuracy of 90%. This is known as the accuracy paradox (Fawcett pg. 862). For a predictive model to be considered effective, it must have a good combination of both successful positive predictions and successful negative predictions (Fawcett pg. 862). Therefore, to overcome the limitations of accuracy, three other criteria are used in this study. They are as follows:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F - measure = \; 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Precision indicates the proportion of the positive predictions that was actually correct, whereas recall measures the proportion of actual positives that were correctly predicted. F-measure provides a single value that considers both precision and recall.

3. Methodology

The data set in this study contained information of over 1700 students and had a total of 18 variables (see Table 4). This information was collected during students' enrollment at Graceland University.

Column Name	Description	Type
Gender Code	Gender of the student.	Nominal
Ethnic Code	Ethnicity of the student.	Nominal
First Generation (Y/N/U)	Indicator of whether the student is first generation or not. Y – yes, N – no, U – unknown.	Nominal
Age at HS Graduation	Students age at high school Graduation.	Discrete
Gap in education	Number gap years between High school graduation and college enrollment.	
Legacy	Indicator of whether student's parents are alumni of the school.	Nominal
ACT/ACT Equivalent SAT Score	ACT or ACT equivalent of the SAT score earned by the student.	Continuous
State	State of which student was a resident of prior to enrolling at Graceland.	Nominal
HS GPA	Student's high school GPA.	Continuous
Cl_Size	Size of the graduating class of which the student was a part of.	Continuous
HS Rank	Rank of the student while graduating from high school.	
New/Transfer	Indicator of whether the student transferred from another institution or not.	Nominal
Declared Major 1	1st major that student had declared during enrollment.	Nominal
Declared Major 2	2nd major that student had declared during enrollment.	Nominal
Sport	Sports team of which student became part of at Graceland.	Nominal
Varsity	Indicator of whether the student played varsity level sport.	Nominal
Family_AGI	Student's family's annual gross income range.	Ordinal

Denomination	Student's religious affiliation.	Nominal
Graduation Status	Indicator of whether the student graduated or not.	Ordinal

Table 4: Data set used in the study.

As shown in Table 4, the data set has four types of variables: nominal, ordinal, discrete and continuous. While nominal variables and ordinal variables are both categorical in nature, the prior simply only has a label without any quantitative value and cannot be ordered whereas the latter is not quantitative yet can be ordered. Continuous and discrete variables, on the other hand, have quantitative value and can be ordered, however, unlike continuous, discrete values only represent integers (Winship and Mare pg. 513).

The main aim of this study is to predict whether a student will graduate or not based on the data collected during enrollment. Therefore, in our data set, the "Graduation Status" is the dependent variable and the rest are independent variables. The independent variables, also known as features will be used to train the machine learning models, which will then predict whether a student will be successfully retained until graduation or not.

Since categorical data don't have any quantitative value, they need to be encoded numerically before they can be used in model training and testing. However, unlike decision tree and naïve Bayes, simply using numerical representations of such values is not enough for logistic regression (Potdar, Pardawala and Pai pg.7). For example, Table 5 contains some sample student resident state data and their respective numerical encodings.

Student	State	Encoding
1	California	1
2	Texas	2
3	Iowa	3

Table 5: Sample resident states and their numerical representation

In Table 5, California has been encoded as 1, Texas as 2, and Iowa as 3. All three state names are simply labels, but the encoding provides them weights. However, as $3 - 2 = 1$, it should also mean that $Iowa - Texas = California$, which is certainly not true.

A common way to encode nominal and ordinal values in machine learning is one-hot encoding. In one-hot encoding, for all N unique values in a column of a data set, additional $N - 1$ columns are created. These columns now contain a value of either 0, denoting the absence of the value and 1, denoting the presence of the value. For example, Table 6 shows one-hot encoding for student data in Table 5.

Student	California	Texas
1	1	0
2	0	1
3	0	0

Table 6: one-hot encoding of student resident state data

In Table 6, the absence of both California and Texas implied that the student was from Iowa, therefore having a separate column for Iowa was redundant. All nominal and ordinal variables in the data set were one-hot encoded using pandas, a data manipulation library for Python before they were used for machine learning. The machine learning models and other statistical tools used in this study were imported from scikit-learn, a machine learning library for Python. Prior to being used for model training and testing, the data set underwent feature engineering and feature selection.

First, the strength of association between each independent variable and the dependent variable was observed. This way the features could be ranked based on their significance. Since the data set was a mixture of both quantitative (continuous and discrete) and qualitative (nominal and ordinal) values, a different approach was taken in order calculate their strength of association with the dependent variable, which itself was qualitative in nature. The correlation between the discrete/continuous independent variables and the dichotomous dependent variable is measured through the Point-Biserial correlation coefficient (Tate pg. 603). Let the dependent variable have two values, 0 (dropped out) and 1 (successfully retained), then, the correlation coefficient is given by the expression:

$$r = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

where M_1 is the mean of the continuous values which corresponds to 1, and M_0 is the mean of the continuous values which correspond to 0, n_1 *and* n_0 are the number of data points in the group of continuous variables that correspond to 1 and 0 respectively, n represents the total number of data points and s_n represents its standard deviation. Similarly, the correlation between two categorical variables is measured by the Cramer's V coefficient (Wu et al. pg. 2595). It is given by the expression:

$$V = \sqrt{\frac{\chi^2}{N * \min(row - 1, col - 1)}}$$

where χ^2 is the chi-squared statistic, N is the number of cases and $\min(row - 1, col - 1)$ represents the minimum of one less the number of row and one less the number of columns of the contingency table created by the two categorical variables being studied. Finally, 5 features were found to have the highest correlation with the "Graduation Status" variable. They are listed in Table 7 in a descending order based on their strength of association with the dependent variable.

Feature	Correlation Coefficient Value
First Generation (Y/N/U)	0.35
High school GPA	0.33
Family_AGI	0.31
ACT/ ACT equivalent SAT score	0.30
State	0.28

Table 7: Top 5 predictors of student retention

It was found that a student's family's educational background had the most influence on their decision to complete college or not. Similarly, student's high school GPA, family's income, aptitude test scores, and their home state had a major influence on the dependent variable.

Two of the three models, i.e. naïve Bayes and logistic regression, compared in this study, assume independence between the features, however, multicollinearity was found between some variables in the data set. Multicollinearity is a state in which one feature can be predicted from other features present in the dataset (Hervé and Williams pg. 433). The features "high school GPA" and "ACT/ ACT equivalent SAT score" were highly correlated with each other with a correlation coefficient of 0.47. In such situation, a common practice is to use principal component analysis (PCA) in order to reduce the dimension of the features involved. When PCA is carried out, a new set of variables are created. These newly created variables are known as principal components and are a linear combination of the original variables (Hervé and Williams pg. 433). However, before applying PCA, the features needed to be scaled. This reduces the margin of error. In this study, standard normal distribution, where the mean is 0 and the standard deviation is 1, was used to scale the variables involved.

The effectiveness of machine learning models is also affected when they are overfitted. Overfitting occurs when too many features are used to train the model and it becomes too complex. After applying one-hot encoding to the categorical variables in the data set, several new variables had been created. For example, the one-hot encoding of the resident state column had created 48 more columns. However, not all these new features are necessary or helpful. Students from some states had more representation than others. For example, only 2 students in the data set were from Vermont. It would be redundant to create an entirely new feature just to represent them. Therefore, certain variables were merged to create new variables. Through trial and error, it was found that grouping states based on which census region they were part of (including U.S. controlled territories) provided the best results. Therefore, 48 columns were reduced to 4. Finally, using the train-test split feature found in the scikit-learn library, 80% of the data was used for training the models, whereas the remaining 20% was used for testing their effectiveness.

4. Results

4.1 Logistic Regression

Table 8 contains the results from logistic regression:

	Actual Positive	Actual Negative
Predicted Positive	139	49
Predicted Negative	31	139

Table 8: Confusion matrix for Logistic Regression

$$Accuracy = \frac{139 + 139}{139 + 139 + 49 + 31} = 0.78 \qquad Precision = \frac{139}{139 + 49} = 0.74$$

$$Recall = \frac{139}{139 + 31} = 0.82 \qquad F - measure = 2 \cdot \frac{0.74 \cdot 0.82}{0.74 + 0.82} = 0.78$$

Logistic regression has an accuracy score of 0.78, precision of 0.74, recall of 0.82 and f-measure of 0.78.

4.2 Naïve Bayes

Table 9 contains the results from logistic regression:

	Actual Positive	Actual Negative
Predicted Positive	140	53
Predicted Negative	42	123

Table 9: Confusion matrix for Naïve Bayes.

$$Accuracy = \frac{140 + 123}{140 + 123 + 42 + 53} = 0.73 \qquad Precision = \frac{140}{140 + 53} = 0.73$$

$$Recall = \frac{140}{140 + 42} = 0.77 \qquad F - measure = 2 \cdot \frac{0.73 \cdot 0.77}{0.73 + 0.77} = 0.75$$

Naive Bayes has an accuracy score of 0.73, precision of 0.73, recall of 0.77 and f-measure of 0.75.

4.3 Decision Tree

Table 10 contains the results from logistic regression:

	Actual Positive	Actual Negative
Predicted Positive	145	40
Predicted Negative	48	125

Table 10: Confusion matrix for Decision Trees

$$Accuracy = \frac{145 + 125}{145 + 125 + 48 + 40} = 0.75 \qquad Precision = \frac{145}{145 + 40} = 0.78$$

$$Recall = \frac{145}{145 + 48} = 0.75 \qquad F - measure = 2 \cdot \frac{0.78 \cdot 0.75}{0.78 + 0.75} = 0.76$$

Decision tree has an accuracy score of 0.75, precision of 0.78, recall of 0.75 and f-measure of 0.76.

4.4 Comparison of Outcomes

	Accuracy (%)	Recall (%)	Precision (%)	F-measure
Logistic Regression	78	82	74	0.78
Naïve Bayes	73	77	73	0.75
Decision Tree	75	75	78	0.76

Table 12: Model effectiveness comparison

As shown in Table 11, logistic regression has the highest f-measure score, followed by decision tree and finally naïve Bayes. Therefore, logistic regression proved to be the most effective model to predict student retention.

5. Conclusion & Future Work

The result of any sort of prediction via machine learning depends on good use of data and algorithms. The correct choice of machine learning method is crucial to achieving the best results, however, it is not sufficient. Feature engineering and feature selection are also an important factor in producing great results.

The aim of this study was to compare the effectiveness of three machine learning models: logistic regression, naïve Bayes and decision trees. Data set of over 1700 students was used to train and test the models. Feature engineering and feature selection process such as PCA was carried out to improve the performance of the models. The evaluation of the models was done through indicators like accuracy, precision, recall, and f-measure. Finally, logistic Regression was found to be the most effective followed by decision tree and naïve Bayes respectively.

This study has certain limitations that need to be noted. The data set used in this study was small. As a result, rows with missing values needed to be imputed to include them into the study. This produces some error in predictions. While working with larger datasets, it would be more affordable to drop rows with missing values. Furthermore, due to government restrictions, access to student disability information was not granted. Knowing whether a student was suffering from any disability or chronic illness could have helped the models make more accurate predictions. Exact family income was also not provided, but instead was provided in terms of income bracket (example: \$50,000 – \$80,000). However, people who belong to the upper range of the bracket don’t experience the same financial stress as those who belong to the lower range. This could have also affected the predictions made by the models. Thus, access to a well-defined data set would offer more conclusive result.

Finally, a variety of machine learning techniques could be researched to get the best results. Similar studies in the past have used random forest and support vector machines to predict student performance in college. Artificial neural networks could also be used to better understand the importance of method selection in producing quality results.

6. References

Abdi, Hervé and Lynne J. Williams. "Principal Component Analysis." *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, 2010, pp. 433-459, http://www.utd.edu/~herve/abdi-awPCA2010.pdf

B. Wu, L. Zhang, and Y. Zhao, "Feature Selection via Cramer's V-Test Discretization for Remote-Sensing Image Classification." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, 2014, pp. 2593-2606, https://ieeexplore.ieee.org/abstract/document/6547700

Fawcett, Tom. "An introduction to ROC Analysis." *Pattern Recognition Letters,* vol 27, no. 8, 2006, pp. 861-874, https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X

Gerdes, Hillary, and Brent Mallinckrodt. "Emotional, Social, and Academic Adjustment of College Students: A Longitudinal Study of Retention." *Journal of Counselling and Development,* vol. 72, no. 3, 1994, pp. 281-288*,* https://psycnet.apa.org/record/1994-23355-001

Ishitani, Terry T. "A Longitudinal Approach to Assessing Behavior among First-Generation Students: Time-Varying Effects of Pre-College Characteristics." *Research in Higher Education*, vol. 44, no. 4, 2003, pp.433-449, http://www.jstor.org/stable/40197314

Maurtaugh, Paul A. et al. "Predicting the Retention of University Students." *Research in Higher Education*, vol. 40, no. 3, 1999, pp 355 – 371, http://science.oregonstate.edu/~murtaugp/files/reshe-1999.pdf

Potdar, Kedar et al. “A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers.” *International Journal of Computer Applications*, vol 175, no. 4, 2017, pp. 7 – 9, https://www.ijcaonline.org/archives/volume175/number4/28474-2017915495

Strobl, Carolin et al. “An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests.” *Psychological Methods,* vol. 14, no. 4, 2009, pp. 323-348, https://psycnet.apa.org/record/2009-22665-002

Tate, Robert F. “Correlation between a Discrete and a Continuous Variable. Point-Biserial Correlation.” *The Annals of Mathematical Statistics*, vol. 25, no. 3, 1954, pp. 603-607, https://www.jstor.org/stable/2236844

Tinto, Vincent. “Research and Practice of Student Retention: What Next?” *J. College Student Retention*, Vol. 8, no. 1, 2006, pp. 1 – 19, http://new.tru.ca/shared/assets/Tinto_2006_what_next23657.pdf

Winship, Christopher and Robert D. Mare. “Regression Models with Ordinal Variables.” *American Sociological Review,* vol. 49, no. 4, 1984, pp. 512-525, https://www.researchgate.net/publication/271685623_Regression_Models_with_Ordinal_Variables

Yan, Jihong and Jay Lee. “Degradation Assessment and Fault Modes Classification Using Logistic Regression.” *ASME. J. Manuf. Sci. Eng.*, vol. 127, no. 4, 2004, pp. 912-914, http://manufacturingscience.asmedigitalcollection.asme.org/article.aspx?articleid=1450204