Predicting Retention Rates from students' Behavior.

Dr. Awatif Amin, Johnson C. Smith University

Awatif Amin is a computer science Assistant professor at Johnson C. Smith University scince 2001. She primarily focuses on programming and data analytics. She completed her Doctorate of Management in organizational Leadership with specialization in Information System Technology (DM/IST), She earned her B.S. and M.S. in Computer Science.

Predicting Retention Rates from students' Behavior.

Abstract

Machine learning and Data mining are supported by the same establishments in two different ways. Machine learning basically learns from present data and delivers the necessary basis for a machine to learn. Data mining uses existing data and finds emerging patterns that help the decision-making processes. Data mining is typically used as an information supplier for machine learning to draw information that recognizes the patterns and determines from these patterns how to adapt behavior for future occurrences. It is obvious to see that there is an overlap between data mining and machine learning as the two have the same goal which is to learn from huge data for analytic resolutions. Machine learning and data mining can be considered knowledge science that concentrates on formulating algorithms that learn from the data and make predictions. Machine learning algorithms include supervised and unsupervised learning classifications. This paper deliberates the use of algorithms to analyze data from educational institutions to help them present more detailed methods to improve the efficiency of recruitment, enrollment, and hence retention.

Introduction

The paper proposed and demonstrates the appropriateness and efficiency of data mining as a method for studying STEM students' retention. The research was to apply machine learning and data mining methods, tools, and algorithms to analyze enrollment data for issues affecting STEM students' retention at Johnson C. Smith University a historically black college. Providing insight into why students drop out before completing their degree, successful identification of students at risk could result in a program of directed retention intercession services. The research question is, what was the relationship between students' commitment behavior, and family background and retention. The approach of this quantitative study was pursuit of an understanding of the factors identified in the literature of retention. The study showed number of class hours, financial situations, lack of family emotional support, social life and institutional assistance were important factors.

Students' retention in higher education has attracted the attention of college and university administrators for many years [1]. According to Bennett, Kottasz, and Nocciolino [2], the retention of college students is a global problem. Retaining students through graduation is an ongoing challenge, costing universities millions of dollars at all levels of higher education, from community colleges to the doctoral level [3]. In the United States, college retention has worsened over several decades, such that "in 1990, the U.S. ranked first in four-year degree attainment among 25-34-year old; [in 2014], the U.S. ranked 12th among other countries" [4]. Data from the National Center for Educational Statistics [5], show many U.S. educational leaders are aware of retention problems in higher education and are making progress in preparing and helping students to raise the retention rate.

In historically black colleges and universities (HBCUs), retention of undergraduate students at four-year institutions has been a long-standing problem and the focus of many past studies. Stakeholders have widely acknowledged that this problem remains complex for policy makers, educators, and college/university administrators [6]. Garland [7] stated that student attrition is a

multidimensional phenomenon involving factors such as financial status, educational background, and family status. Compounding the problem is the minimum participation of HBCU students in science and technology, which is a vital issue [8]. Science, technology, engineering, and math (STEM) minority students represent very few resources [9]. The absence of minority students in the STEM field of education and in the workforce, is one of the foremost contributors to STEM personnel shortages in the U.S. [9]. In 2015, minorities constituted about 45% of baccalaureate graduates and 40% of all STEM college degrees awarded. African Americans accounted for only 8.7% of all 4-year degrees in STEM fields [10]

This quantitative, descriptive, and retrospective study involved data mining methods, tools, and algorithms to sift through enrollment data to identify and analyze issues affecting students' retention. Data consisted of students' demographics, background, behavior, and persistence relevant to enhancing retention among students who express unusual potential in all the fields related to science, technology, engineering and math. Students' enrollment data from a minority university were useful data sources to identify students most likely to leave the institution prematurely. Successful identification of "at risk" students could result in targeted retention intervention services. Although using data mining tools is new to higher education, the approach is widespread in many industries including the intelligence communities and business settings to predict a range of customer behaviors, including customer attrition [11]. Analysts apply data mining to numerous different applications, such as summarizing data, analyzing changes, learning classification rules, finding relations, and detecting inconsistencies [12].

Problem Background

The number of U.S. students obtaining a bachelor's degree in STEM and related fields is inadequate to meet the demand for scientists and engineers [13]. Thus, U.S. businesses are exporting jobs to other countries. Importing skilled personnel, exporting jobs, and increasing H-1B visa allotments do not constitute sound national policy [14]. Understanding decreasing retention among HBCU minority students in general and specifically in the STEM area of study is vital. According to Chubn, May, and Babco [15], African American and Latino students who earned bachelor's degrees in engineering were 4.6% and 6.2%, respectively. Passel, Livingston and Cohn et al. stated [16] that for the United States to compete globally, the nation needs skilled STEM graduates for sustained economic growth and global effectiveness.
The importance of the research is that retention of STEM students has gained the attention of higher education leaders throughout the country, including HBCU institutions [17]. Faculty and administrators of HBCUs should make every effort to continue to retain their STEM students who can do satisfactory academic work despite obstacles such as lack of resources and funding. Despite the important contributions of HBCUs to U.S. economic growth, these institutions receive insufficient support. Suitts [18] wrote, "during the 1990s, for instance, HBCUs received less than 2% of the total amount of $140 billion in federal grants awarded to America's institutions of higher education for science and engineering programs" (p. 205).

Research is sparse and the body of knowledge small regarding the effectiveness of data mining algorithms applied to the student retention problem. A data mining approach could be a most worthwhile strategy for other practitioners and researchers planning to include many variables

along with all levels of STEM students in the data set. Hendrix [19] found only one out of every 2,352 dissertation abstracts included the search words "data mining" in a search of dissertations with the key words "higher education" and "retention."

In this review, the relevant retention factors, organized in groups, concern students' academic background, commitment behavior, and family background. Multiple variables within these three groups of factors are also part of the discussions.

Method & Model

The data mining method of analysis was a predictive classification process using decision trees for model development. In machine learning, analysts refer to prediction methods as supervised learning methods [20]. Models usually describe and explain facts that had been unknown or buried in the data set, and conversely, show the structure of the newly discovered relationship between the independent variables [21]. These facts are useful for predicting the importance of the dependent variable when the independent variables are available.

Retention prediction by commitment behavior attributes and first semester and first year GPA.

Table 1 contains descriptive statistics concerning commitment to degree attainment for 101 students studied. The dependent attribute variable was degree/no degree. The independent attributes were hours attempted and earned, withdrawal, and grade point average for the first semester and the first year; also, first year STEM GPA (see Figure 1 and Table 2).

Table 1
Descriptive Statistics of 101 Students' Commitment Behavior in the First College Year

ID	*Type*			*Average*
Dependent Variable Degree/No Degree	Categorical	No Degree 49	Degree 52	Total 101
Independent Variables:		*Min*	*Max*	*Average*
Semester Hours Attempted	Integer	9	18	15.455
Semester Hours Earned	Integer	0	18	14.703
Semester Withdrawal	Integer	0	16	0.812
Semester GPA	Real	0	4	2.922
First Year Hours Attempted	Integer	0	40	28.673
First Year Hours Earned	Integer	0	40	27.822
First Year Withdrawal	Integer	0	19	1.260
First Year GPA	Real	0	4	2.742
First Year STEM GPA	Real	0	4	2.255

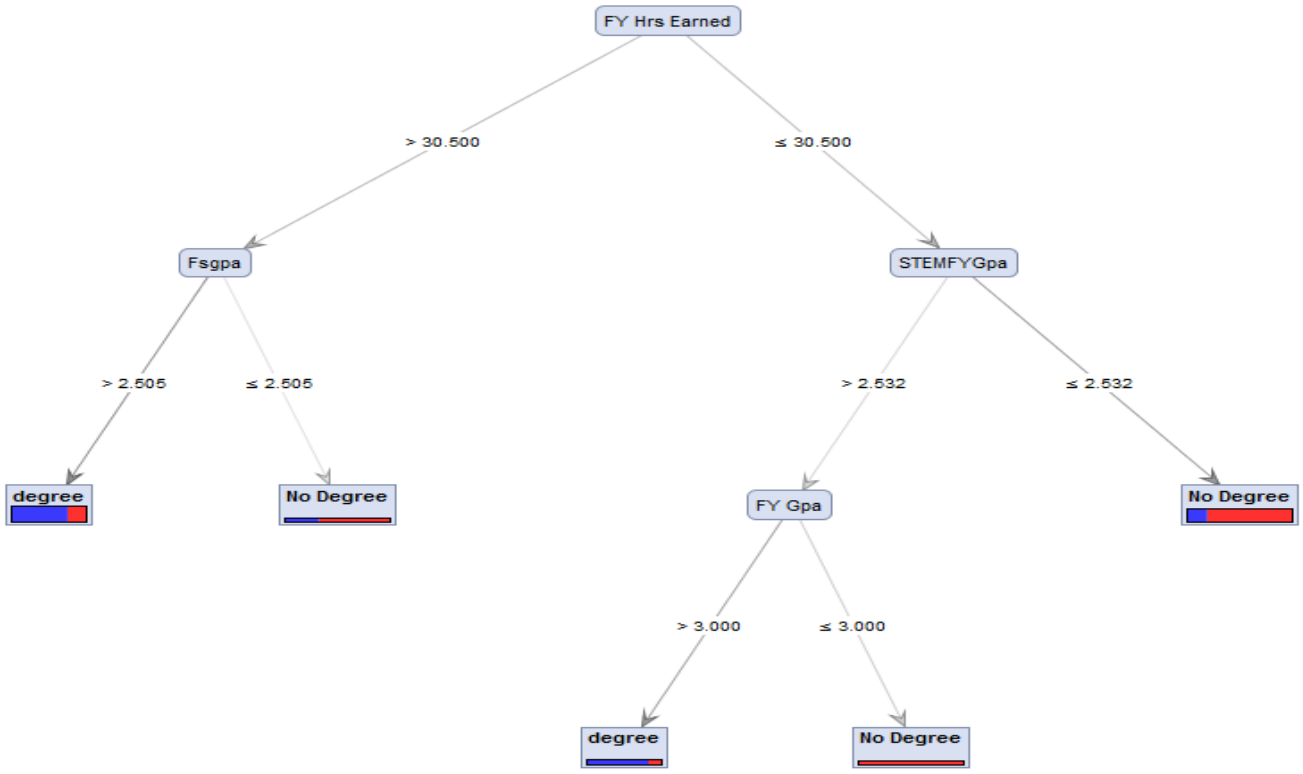

Figure 1. Retention decision tree of commitment behavior of 101 students. FY = First year, Fsgpa = First semester grade point average, STEMFYGpa = First year STEM GPA. STEM fields are science, technology, engineering, and mathematics.

Table 2

Retention of Students by Commitment Behavior and GPA: Narrative Description

First Year Hours Earned > 30.500

| First Semester GPA > 2.505: degree {degree=38, No Degree=12}
| First Semester GPA ≤ 2.505: No Degree {degree=1, No Degree=2}

First Year Hours Earned ≤ 30.500

| STEM First Year GPA > 2.532
| | First Year GPA > 3.000: degree {degree=5, No Degree=1}
| | First Year GPA ≤ 3.000: No Degree {degree=0, No Degree=2}
| STEM First Year GPA ≤ 2.532: No Degree {degree=8, No Degree=32}

Note: GPA = Grade point average.

For the class *no degree*, the confusion matrix in Table 11 shows the accuracy rate or effectiveness of the model was .60 or 60%, calculated as (13+5)/ (13+5+7+5). The precision rate of .722 for the class *no degree* indicates that 72.2 % of those predicted to leave with no degree actually got no degree, calculated as 13/(13+5). Calculation of the true positive rate, termed

recall rate for the class *no degree* was .65, calculated as (13/(13+7), indicating successful identification of the 65% of students who did obtain a degree.

Table 3

Confusion Matrix for Commitment Behavior Decision and GPA Tree

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 5	False Positive (FP) = 7
No Degree	False Negative (FN)= 5	True Negative (TN) = 13

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction.
*Accuracy % = (TP + TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

To determine if there is a relationship between students' commitment behavior, first semester GPA, and first year GPA and retention, χ^2_{+p} and χ^2_{+r} need to be calculated.

χ^2_{+p} = ((TP-ETP)^2/ETP+(FP-EFP)^2/EFP)

= ((5- 4) ^{2}/4) + (7-7.2) ^{2}/7.2))

= (0.25+ 0.2)

= 0.45

χ^2_{+r} = ((FN -EFN)^2/EFN+(TN-ETN)^2/ETN)

= ((5-6) ^{2}/6) + (13- 12) ^{2}/12))

= (0.16667 + 0.083)

= 0.1749

Since $\chi^2{}_{(\alpha)}$= 3. 841459 and both χ^2_{+p} and $\chi^2_{+r} <$ 3.841459, the relationship between students' commitment behavior, first semester GPA, and first year GPA and retention is insignificant although from the confusion matrix the cell frequency >5.

However, the decision tree result in Figure 1 and Table 1 shows that if first-year hours earned were > 30.5 and the first semester GPA was > 2.5, 38 out of 50 students obtained a degree (76%). If first-year hours earned were <= 30, with a STEM GPA of > 2.5, and first year GPA of >3.0, 5 out of 6 students (83%) received a degree. The tree also shows if first-year hours earned were > 30.5 and STEM GPA was < = 2.5, there was a 20% lower chance, 8 out of 40, that a student would receive a degree.

For the class *no degree*, the confusion matrix in Table 3 shows the accuracy rate or effectiveness of the model was .60 or 60%, calculated as (13+5)/ (13+5+7+5). The precision rate of .722 for the class *no degree* indicates that 72.2 % of those predicted to leave with no degree actually got no degree, calculated as 13/(13+5). Calculation of the true positive rate, termed *recall* rate for the class *no degree* was .65, calculated as (13/ (13+7), indicating successful identification of the 65% of students who did obtain a degree.

Retention prediction by commitment behavior, second iteration: Without GPA data.

Table 4 contains descriptive statistics for a second iteration of the decision tree of independent attributes concerning commitment to degree attainment with omission of First Year GPA and First Year STEM GPA. The dependent attribute variable remained degree/no degree. The independent attributes were hours attempted and earned, and withdrawal, for the first semester and the first year. Figure 2 and Table 5 show the results of the second iteration decision tree in graphic and narrative form, respectively, concerning commitment behavior attributes with first semester GPA and first year GPA omitted.

Table 4
Descriptive Statistics of Commitment Behavior—Second Iteration without GPA Data

ID	*Type*			*Average*
Dependent Variable		No Degree	Degree	Total
Degree/No Degree	Categorical	49	52	101
Independent Variables:		*Min*	*Max*	*Average*
First Semester Hours Attempted	Integer	9	18	15.455
First Semester Hours Earned	Integer	0	18	14.703
First Semester Withdrawal	Integer	0	16	0.812
First Year Hours Attempted	Integer	0	40	28.673
First Year Hours Earned	Integer	0	40	27.822
First Year Withdrawal	Integer	0	19	1.260

Figure 2 and Table 5 show the results of the second iteration decision tree in graphic and narrative form, respectively, concerning commitment behavior attributes with first semester GPA and first year GPA omitted.

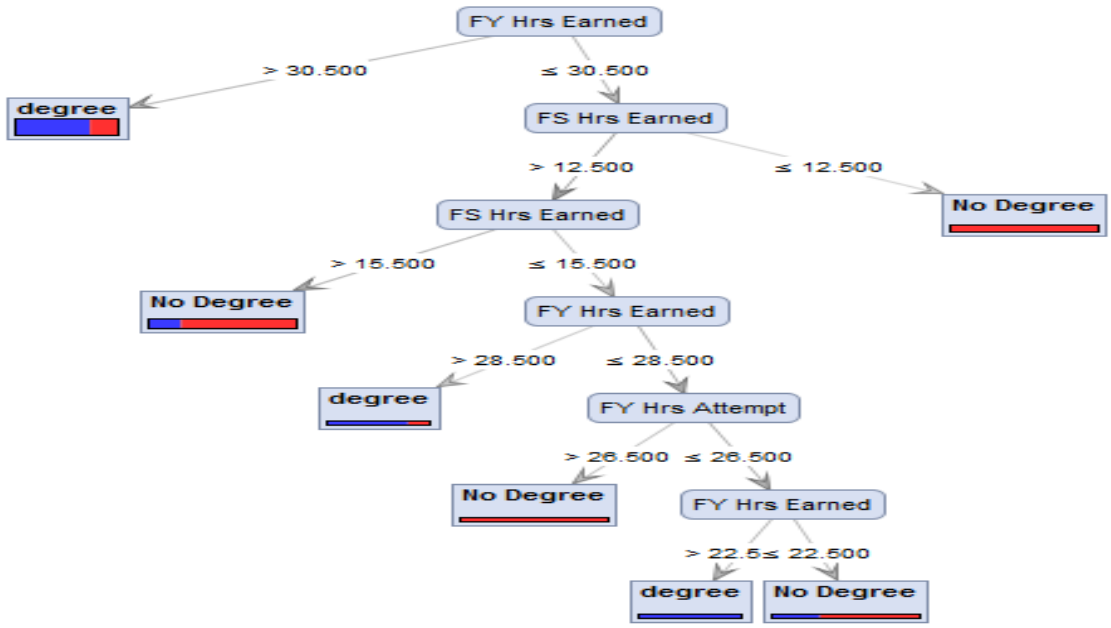

Figure 2. Commitment behavior decision tree, second iteration is without GPA and STEM GPA data. FS = first semester; FY = first year.

Table 5

Retention Prediction by Commitment Behavior: Narrative Description of Decision Tree, Second Iteration without GPA

FY Hrs Earned > 30.500: degree {degree=39, No Degree=14}
FY Hrs Earned ≤ 30.500
| FS Hrs Earned > 12.500
| | FS Hrs Earned > 15.500: No Degree {degree=5, No Degree=17}
| | FS Hrs Earned ≤ 15.500
| | | FY Hrs Earned > 28.500: degree {degree=4, No Degree=1}
| | | FY Hrs Earned ≤ 28.500
| | | | FY Hrs Attempt > 26.500: No Degree {degree=0, No Degree=4}
| | | | FY Hrs Attempt ≤ 26.500
| | | | | FY Hrs Earned > 22.500: degree {degree=3, No Degree=0}
| | | | | FY Hrs Earned ≤ 22.500: No Degree {degree=1, No Degree=2}
| FS Hrs Earned ≤ 12.500: No Degree {degree=0, No Degree=11}

Note: FS = First Semester; FY = First Year.

For the class *no degree*, the confusion matrix in Table 6 shows the accuracy rate or effectiveness of the model was .56 or 56%, calculated as (6+11)/ (6+4+9+11). The precision rate of .733 shown for the class *no degree* indicates that 73.3 % of those predicted to leave with no degree actually did not obtain a degree, calculated as 11/(11+54). Calculation of the true positive rate, termed *recall* rate for class *no degree* was (11/(11+9) = .55, indicating successful identification of the 55% of students who obtained a degree.

Table 6
*Confusion Matrix for Commitment Behavior Decision Tree, No GPA: *Accuracy = 56%, ** Precision = 73.33%, ***Recall = 55% (Positive Class = No Degree)*

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 6	False Positive (FP) = 9
No Degree	False Negative (FN)= 4	True Negative (TN) = 11

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction.
*Accuracy % = (TP + TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

To determine if there is a relationship between students' commitment behavior and retention, $\chi^2_{+p} + \chi^2_{+r}$ needs to be calculated.

$\chi^2 = + \chi^2_{+r}$

$\chi^2_{+p} = ((TP-ETP)^2/ETP+(FP-EFP)^2/EFP)$

$= ((6- 5)^2/5) + (9-10)^2/10))$

$= (0.20+ 0.1)$

$= 0.3$

$\chi^2_{+r} = ((FN -EFN)^2/EFN+(TN-ETN)^2/ETN)$

$= ((4-5)^2/5) + (11- 10)^2/10))$

$= (0.2 + 0.1) = 0.3$

Since $\chi^2_{(\alpha)} = 3.841459$ and both χ^2_{+p} and $\chi^2_{+r} < 3.841459$, the relationship between students' commitment behavior and retention is insignificant although from the confusion matrix the cell frequency <5. However, the result of the decision tree in Figure 13 and Table 13 show if first year hours earned were > 30.5, 39 out of 53 students received a degree (74%). If first year hours earned were <= 30.5 and first semester hours earned were > 15.5, only 5 out of 22 received a degree (23%). When considering slightly fewer first year hours earned, > 28.5 but <= 30.5 and <= 15.5 first semester hours earned, 4 out of 5 students received a degree (80%). The tree also shows if first semester hours earned were <= 15.5, and first year hours attempted were between 22 and 26, 3 out of 3 students (100%) received a degree. If first year hours earned were <= 30.5 and first semester hours earned were <= 12, none obtained a degree.

Family background and emotional support/social attributes.

Table 7 contains descriptive statistics concerning 101 students' attributes, in which the dependent variable was degree/no degree, and the independent attributes were aspects of income, student's social and emotional support resources, employment status, and parent's education. Family average gross income or AGI and total amount of financial aid (Tot Aid) ranked from lowest to highest. Variables ranked 1 to 10, with 10 highest were: Discuss family problems with

counselor (DFPC), family emotional support (FES), receptivity to institutional assistance (RtIA), receptivity to social enrichment (RtSL), sense of financial security (SFS), get help with study (HS), study habits (SH), and school athletics team member (AthTM), Mother's and fathers' highest education grade level (MHGL, FHGL), ranked 0 to 3, where zero was unknown, 1 denoted middle school, 2 denoted high school, and 3 was college or above. Job work load ranks were 0-None, 1 = 1-10 hours/week, 2 = 11-20 hours/week, and 3 = 20-40 hours/week.

Table 7
Descriptive Statistics of Students' Family Background and Emotional/Social Attributes

ID--Dummy Code	*Type*				*Average*
Dependent Variable			No Degree	Degree	Total
Degree/No Degree	Categorical		49	52	101
Independent Variables:			*Min*	*Max*	*Average*
Family Average Gross Income (AGI)	Integer		0	175,447	35,960
Total Financial Aid (TotAid)	Integer		0	30,220	20,217
Discuss Family Problems with Counselor (DFPC)	Real		0	9.31	4.13
Family Emotional Support (FES)	Real		0	26	4.06
Receptivity to Institutional Assistance (RtIA)	Real		0	9.9	4.96
Receptivity to Social Enrichment (RtSL)	Real		0	9.9	4.29
Sense of Financial Security (SFS)	Real		0	9.9	3.64
Mother's Highest Education Grade Level (MHGL)	Integer		0	3*	2.07
Father's Highest Education Grade Level (FHGL)	Integer		0	3*	1.54
Job Work Load (JWL)	Integer		0	3**	1.00
					Values
Study Habits (SH)	Polynomial		9.9 (1)	0 (26)	0 (26) 3.6 (9) [24 more]
Get Help with Studies (HS)	Polynomial	0	9.51	0 (26)	0 (26), *(1), [74 more]
School Athletics Team Member (AthTM)	Polynomial	0	Yes (11)	No (90)	N(90) Y(11)

**Note.* *Mother's and Father's Highest Education Level ranked from 0 to 3 as follows: 0 = unknown, 1 = middle school, 2 = high school, and 3 = college and above. **Job Work Load rankings were from 0 to 3 as follows: 0-None, 1 = 1-10 hours/week, 2 = 11-20 hours/week, and 3 = 20-40 hours/week.

Decision tree of finances and emotional/social attributes (see Figure 3 and Table 8).

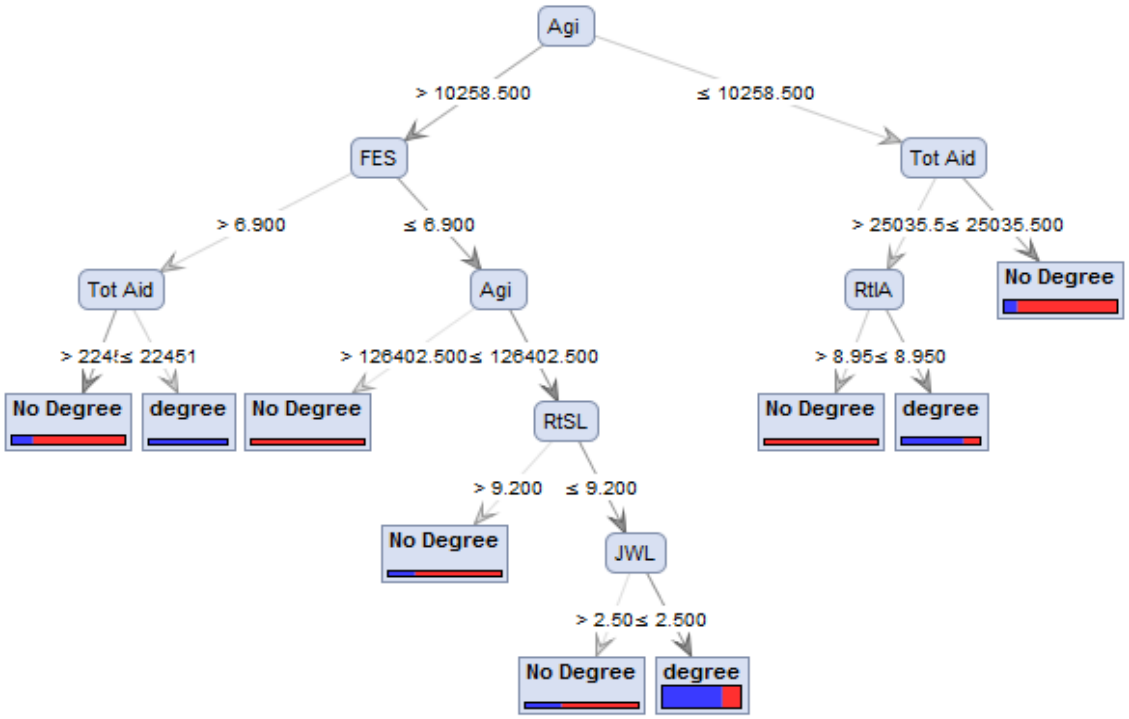

Figure 3. Family background and emotional support/social aspect decision tree for retention contains economic, emotional, and social attributes relevant to retention. Agi = family adjusted gross income, Tot Aid = financial aid, FES = family emotional support, RtIa = receptivity to institutional assistance, RtSL = receptivity to social enrichment, and JWL = job work load.

Table 8
Retention Prediction by Family Background and Emotional/Social Attributes: Narrative Description of Decision Tree, First Iteration

Agi > 10258.500
| FES > 6.900
| | Tot Aid > 22451: No Degree {degree=2, No Degree=8}
| | Tot Aid ≤ 22451: degree {degree=3, No Degree=0}
| FES ≤ 6.900
| | Agi > 126402.500: No Degree {degree=0, No Degree=2}
| | Agi ≤ 126402.500
| | | RtSL > 9.200: No Degree {degree=1, No Degree=3}
| | | RtSL ≤ 9.200
| | | | JWL > 2.500: No Degree {degree=1, No Degree=2}
| | | | JWL ≤ 2.500: degree {degree=38, No Degree=11}
Agi ≤ 10258.500
| Tot Aid > 25035.500
| | RtIA > 8.950: No Degree {degree=0, No Degree=2}
| | RtIA ≤ 8.950: degree {degree=4, No Degree=1}
| Tot Aid ≤ 25035.500: No Degree {degree=3, No Degree=20}

Note: FS = First Semester; FY = First Year.

Confusion matrix of finances and emotional/social attributes. For the class *no degree*, the confusion matrix in Table 9 shows the tree produced an accuracy rate of .53 or 53% effectiveness of the model calculated as (8+8)/ (8+2+12+8). The precision rate was .80 calculated as 8/(8+2) indicated 80% of those predicted to leave without a degree actually did not receive a degree. The recall rate was .40 indicating successful identification of the 40% of students who did obtain a degree (8/ (8+12) *100 = 40%).

Table 9
*Confusion Matrix for Family Income and Emotional/Social Attributes Decision Tree: *Accuracy = 53%, ** Precision = 80%, ***Recall = 40% (Positive Class = No Degree)*

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 8	False Positive (FP) = 12
No Degree	False Negative (FN)= 2	True Negative (TN) = 8

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction.
*Accuracy % = (TP + TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

To determine if there is a relationship between family background and emotional support/social aspect and retention, χ^2_{+p} *and* χ^2_{+r} need to be calculated.

χ^2_{+p} = ((TP-ETP)^2/ETP+(FP-EFP)^2/EFP)

= ((8-6.67)^2/6.667) + (12-6.667)^2/6.67))

= (0.266+ 4.26)

= 4.526

χ^2_{+r} = ((FN -EFN)^2/EFN+(TN-ETN)^2/ETN)

= ((2-3.333)^2/3.33) + (8 – 6.67)^2/6.67))

= (0.53 + .365)

= 0.895

Since $\chi^2_{(\alpha)}$= 3. 841459

and$\chi^2_{+p} > 3.841459$, the relationship between students' commitment behavior and retention is significant although from the confusion matrix the cell frequency <5.

Figure 3 decision tree and Table 8 show the results of the first iteration decision tree in graphic and narrative form, respectively, concerning finances and emotional/social aspects of family background and student attributes. One relationship was between family emotional support

(FES) and finances. If the rank of family emotional support was seven or above, income was greater than $10,000 per year, and students' total aid was greater than $22,000 per year, only 25% of students received a degree, but 100% of students with financial aid less than $22,000 received a degree. If family emotional support was lower than 7 and family income was above $126,000 per year, none received a degree. For students ranked nine or higher out of 10 for receptivity to social enrichment (RtSL) with family income less than $120,000, only 25% received a degree. In contrast, students ranked less than nine for RtSL with a job load less than 30hrs/week, 77.7 % received a degree. For students with a job load over 30hrs/week, only 50% received a degree. Receptivity to institutional assistance (RtIA) was another attribute of interest for its potential relationship to degree attainment. The tree shows among students ranking less than eight for RtIA, with family income less than $10,000 and receiving financial aid over $25,000, 80% received a degree. When RtIA was above eight, none received a degree. Students with family income less than $10,000 and receiving financial aid less than $25,000, only 13% received a degree.

Family background and support second iteration decision tree. Table 7 contains descriptive statistics concerning 101 students' attributes, in which the dependent variable was degree/no degree, and the independent attributes were aspects of income, student's social and emotional support resources, employment status, and parent's education. Family average gross income or AGI and total amount of financial aid (Tot Aid) ranked from lowest to highest. Variables ranked 1 to 10, with 10 highest were: Discuss family problems with counselor (DFPC), family emotional support (FES), sense of financial security (SFS), mother's and fathers' highest education grade level (MHGL, FHGL) were ranked from 0 to 3, where zero was unknown, 1 denoted middle school, 2 denoted high school, and 3 was college or above. Job work load (JWL) ranks were 0-None, 1 = 1-10 hours/week, 2 = 11-20 hours/week, and 3 = 20-40 hours/week.

Figure 4 and Table 11 show the results, in graphic and narrative form respectively, of a decision tree applied to three of the attributes concerning family income and student emotional/social condition. The dependent attribute remained degree/no degree. The independent attributes were family adjusted gross income (AGI), discuss family problems with counselor (DFPC), and mother's highest education grade level (MHGL). See Table 10 for the descriptive statistics of the selected attributes.

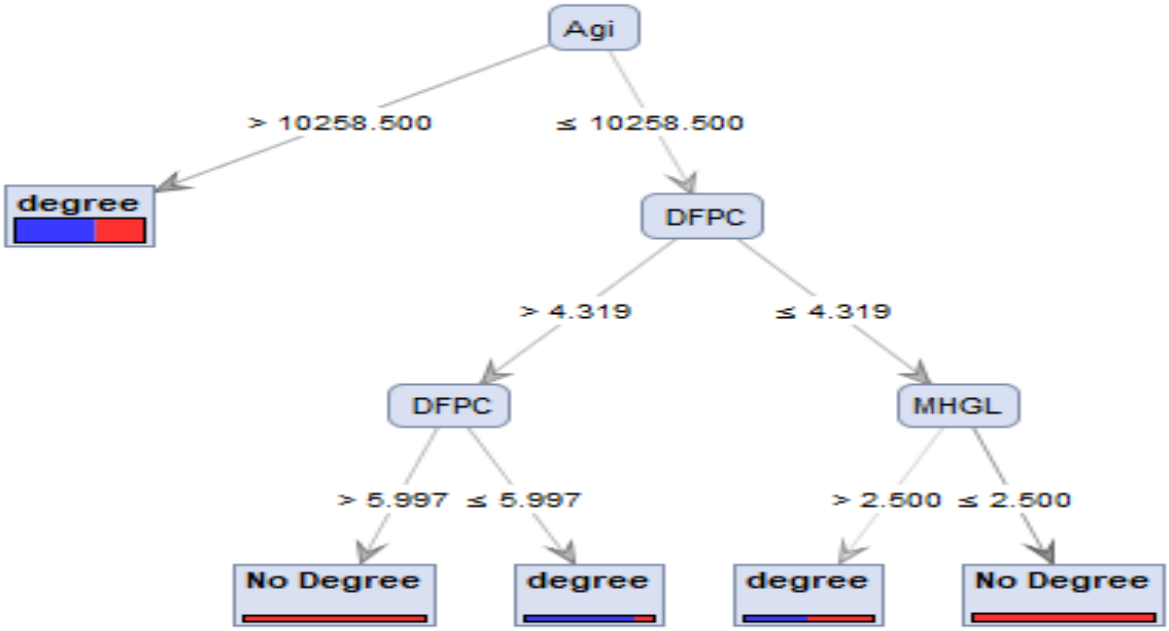

Figure 4. Attributes of the parents income-mother's education-counselor contact decision tree were AGI = family adjusted gross income, DFPC = Discussed family problem with counselor, and MHGL = Mother's highest grade level.

Table 11

Family Background and Emotional/Social Support: Second Iteration

Agi > 10258.500: degree {degree=45, No Degree=26}
Agi ≤ 10258.500
| DFPC > 4.319
| | DFPC > 5.997: No Degree {degree=0, No Degree=7}
| DFPC ≤ 4.319
| | MHGL > 2.500: degree {degree=1, No Degree=1}
| | MHGL ≤ 2.500: No Degree {degree=0, No Degree=14}
Agi > 10258.500: degree {degree=45, No Degree=26}
Agi ≤ 10258.500
| DFPC > 4.319
| | DFPC > 5.997: No Degree {degree=0, No Degree=7}
| DFPC ≤ 4.319

Note: AGI = Family adjusted gross income, DFPC = Discuss family problems with counselor, MHGL = Mother's highest grade level.

Confusion matrix for Family Background and Emotional/Social Support: Decision tree second iteration in Figure 4. For the class *no degree*, the confusion matrix in Table 12 shows the accuracy rate or effectiveness of the model was .466 or 46.6%, calculated as (9+5)/ (9+1+15+5). The precision rate of .833 shown for the class *no degree* indicates that 83.3% of those predicted to withdraw from the university with no degree actually did not obtain a degree, calculated as 5/(1+5). Calculation of the true positive or recall rate for the class *no degree* was 5/(5+15) =.25, indicating successful identification of the 25% of students who received a degree.

Table 12
Confusion Matrix for Family Background and Emotional/Social Support Decision Tree

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 9	False Positive (FP) = 15
No Degree	False Negative (FN)= 1	True Negative (TN) = 5

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction.
*Accuracy % = (TP + TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

To determine if there is a relationship between students' parents' income, parents' education, and counselor contact and retention, χ^2_{+p} and χ^2_{+r} need to be calculated.

χ^2_{+p} = ((TP-ETP)^2/ETP+(FP-EFP)^2/EFP)

= ((9- 8) ^{2}/8) + (15-16) ^{2}/16))

= (0.125+ 0.06)

= 0.185

χ^2_{+r} = ((FN -EFN)^2/EFN+(TN-ETN)^2/ETN)

= ((1-2) ^{2}/2) + (5- 4) ^{2}/4))

= (0.5 + 0.25)

= 0.75

Since $\chi^2{}_{(\alpha)}$= 3. 841459, and both χ^2_{+p} and $\chi^2_{+r} <$ 3.841459, the relationship between students' parents' income, parents' education, and counselor contact and retention is insignificant, from the confusion matrix the cell frequency <5.

Figure 4 decision tree and Table 11 show when family income was above $10,000 per year, 63% of students received a degree. No students received a degree having family income less than $10,000 per year and discussed family problems with a counselor (DFPC) at above-average frequency. Almost 50% of students did receive a degree having the same income status, where family income < $10,000.00 per year, who discussed family problems with counselor (DFPC) at a below-average rate and with a mother's highest education grade level (MHGL) above high school. However, no students received a degree having the same financial and social conditions, with a mother's highest education grade level (MHGL) less than high school graduate or some college.

Institutional assistance-social life, study habits, and help decision tree. The dependent variable for the decision tree remained if the student received a degree or no degree. The independent attributes concerned students' academic and extracurricular activities. Attributes were receptivity to institutional assistance (RtIA), receptivity to social enrichment (RtSL), get help with study (HS), study habits (SH), and school athletics team membership (AthTM). Table

7 contains descriptive statistics of the attributes of the decision tree; Figure 5 and Table 12 contain the tree and the narrative form respectively.

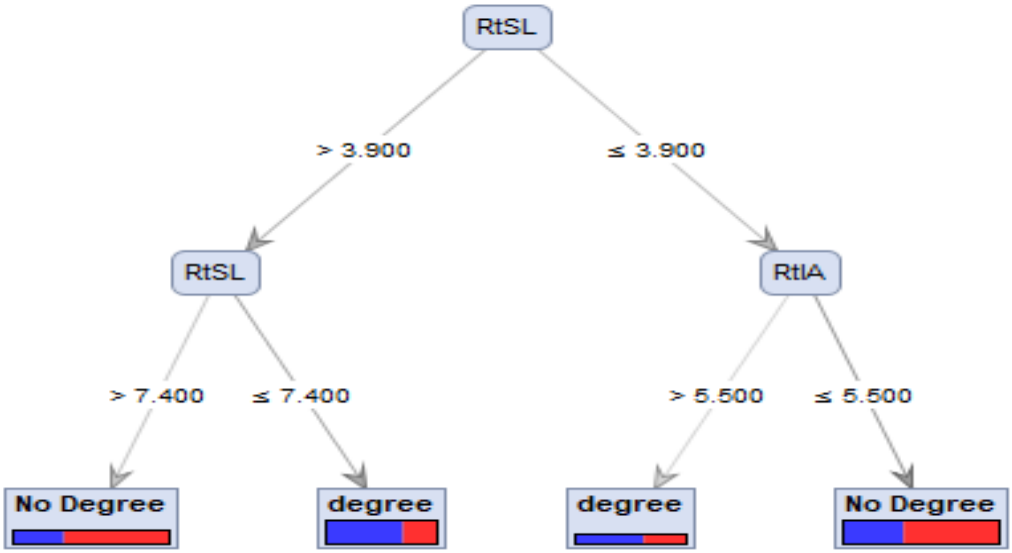

Figure 5. The decision tree representing retention factors consisted of independent attributes of students' academic and extracurricular activities. RtIA = receptivity to institutional assistance and RtSL = receptivity to social enrichment.

Table 12

Narrative Description of Assistance-Social Life-Athletics-Study Habits Decision Tree

RtSL > 3.900
| RtSL > 7.400: No Degree {degree=6, No Degree=12}
| RtSL ≤ 7.400: degree {degree=26, No Degree=11}
RtSL ≤ 3.900
| RtIA > 5.500: degree {degree=5, No Degree=3}
| RtIA ≤ 5.500: No Degree {degree=15, No Degree=23}
RtSL > 3.900
| RtSL > 7.400: No Degree {degree=6, No Degree=12}
| RtSL ≤ 7.400: degree {degree=26, No Degree=11}
RtSL ≤ 3.900
| RtIA > 5.500: degree {degree=5, No Degree=3}
| RtIA ≤ 5.500: No Degree {degree=15, No Degree=23}

Note: RtSL = Receptivity to social enrichment/social life. RtIa = Receptivity to institutional assistance.

Confusion matrix for institutional assistance-social life decision tree**.** For the class *no degree*, the confusion matrix in Table 13 shows the accuracy rate was .3667 or effectiveness of the model was 36.67%, calculated as (8+3)/(8+2+17+3). The precision rate of .60 for the class *no degree* indicates that 60% of those predicted to withdraw from the university with no degree actually did not receive a degree, calculated as 8/(8+17). Calculation of the true positive or

recall rate for the class *no degree* was 3/(3+17) =.15, indicating successful identification of the 15% of students who received a degree.

Table 13

*Confusion Matrix for Institutional Assistance-Social Life Decision Tree: *Accuracy = 36.67%, ** Precision = 60%, ***Recall = 15% (Positive Class is No Degree)*

Model Prediction	Actual Degree/No Degree Status from Test Set	
	Degree is True	No Degree is True
Degree	True Positive (TP) = 8	False Positive (FP) = 17
No Degree	False Negative (FN)= 2	True Negative (TN) = 3

Note: TP = correct positive prediction; FP = incorrect positive prediction; FN = incorrect negative prediction. TN = correct negative prediction. *Accuracy % = (TP+TN)/total number of students (TP+FN+FP+TN). **Precision = TP/total predicted positives (TP+FP). ***Recall or true positive rate = TP/total actual positives (TP+FN).

To determine if there is a relationship between students' academic and extracurricular activities and retention, χ^2_{+p} and χ^2_{+r}need to be calculated.

χ^2_{+p} $= (\text{TP-ETP})^2/\text{ETP}) + (\text{FP-EFP})^2/\text{EFP})$

$= ((8- 8.333)^2/8) + (17-16.67)^2/16.67))$

$= (0.0139+ 0.007)$

χ^2_{+r} $= ((\text{FN -EFN})^2/\text{EFN}+(\text{TN-ETN})^2/\text{ETN})$

$= ((2-1.667)^2/1.667) + (3- 3.33)^2/3.33))$

$= (0.067+0.032)$

$= 0.099$

Since $\chi^2_{(\alpha)} = 3.841459$,

and both χ^2_{+p} and $\chi^2_{+r} < 3.841459$, the relationship between students' academic and extracurricular activities and retention is insignificant, from the confusion matrix the cell frequency <5.

However, figure 16 and table 19 show that relatively low social enrichment and higher receptivity to institutional assistance improved degree attainment. When students' receptivity to social enrichment (RtSL) was above 7.4 on the 10-point scale with 10 as the maximum, only 50% received a degree, in contrast to the 70% of students receiving a degree with an RtSL level below 7.4. If RtSL was less than 3.9, and receptivity to institutional assistance (RtIA) was above 5.5, 65.5% received a degree; when RtIA was less than 5.5, only 39.4 % received a degree.

Results

The research shows the balance between the number of first-year credit hours, first semester GPA, and first year GPA is very important.

- Nearly all the students received degrees who earned 30 credit hours or more the first year with greater than a 2.5 STEM GPA in the first semester. Of the students who earned more than 30 credit hours the first year but had less than a 2.5 STEM GPA in the first semester, few successfully completed the degree program.
- Nearly all students obtained degrees who earned fewer than 30 credit hours in the first year, if they achieved a first semester STEM GPA greater than 2.5 and a general GPA of 3.0 or more.
- For students that earned less than15 hours in the first semester, the prediction indicated their chance of surviving increased by taking 22 to 26 credit hours per year.

The research also shows the importance of family emotional support and receptivity to institutional assistance to degree attainment, in contrast to a high level of receptivity to social enrichment.

- For students with high family emotional support and their parents' income above $126,000 per year and they got a large amount of aid not many graduated, where students who got less aid all graduated.
- Almost all students graduated whose parents had relatively less money, got large amounts of financial aid, and showed an average level of receptivity to institutional assistance.
- Students with middle class and high-class parents' income and low family emotional support successfully completed a degree program. However, few students obtained a degree whose parents' income was average and their receptivity to social enrichment was high. Nearly all students completed a degree program who had low receptivity to social enrichment and worked 20 to 30 hours per week.
- Few students received degrees whose receptivity to institutional assistance was high.
- Students did not receive degrees who were from low income families and who discussed family problems with counselors at above-average to high rates.
- Half the students received degrees who discussed family problems with a counselor at below-average rates and had mothers educated above high school.

Conclusion

This paper represents the student retention at Johnson C. Smith University, a historically black higher education institution (HBCU), data mining was the method used to predict why students withdraw before achieving a college degree. Persevering until graduation was contingent upon enough financial resources, institutional assistance, academic support, student's low or moderate need to work, family emotional support, and a limited social life. The multiple factors affecting retention require solutions from the leadership and administration at the university, including approaches to explore, reinforce, or reform fiscal programs that address the entirety of students' financial needs relevant to successful completion of their academic goal. A partnership between leaders, faculty, and alumni could create a strategic plan of action in a centralized center for meeting students' needs to solve the retention problem at Johnson C, Smith University.

Although this paper used the data from one private urban HBCU, continued research is needed to evaluate the impact of retention factors on larger private and public institutions. Machine Learning and data mining can be very rewarding as researchers can apply many different methods to institutions of all sizes

and types as needed. The suggestion of establishing a centralized center supporting different kinds of research to solve retention problems could impact the university's marketing and recruitment activities as well. Improved management of new, innovative, and existing resources could improve retention and allow for greater financial stability at Jonson C. Smith University.

References

[1] L. A. Spakman, W. S. Maulding and j. G. Roberts, "Non-cognitive predictors of student success in college.," *College Student Journa,* p. 46, Fall 2012.

[2] R. Bennett, ,. R. Kortaz and j. Nocciolino, "Catching the early walker: An examination of potential antecedents of rapid student exit from business-related undergraduate degree programs in a post-1992 university.," *Journal of Further and Higher Education,* vol. 2, no. 31, pp. 109-132, 2012.

[3] E. Statistics., "Unergraduate Retention and Graduation Rates," in . *The Condition of Education 2017 (NCES 2017)*, (2017).

[4] Higher-Education., "Higher education—The White House.," (2014). [Online]. Available: http://www.whitehouse.gov/issues/education/higher-education.

[5] C. Demetriou and A. Schmitz-Schmitz, "Integration, motivation, strengths and optimism: Retention theories past, present and future. In R. Hayes (Ed.),," in *Proceedings of the 7th national symposium on student retention* , Norman, 2011.

[6] M. Garlang, "Student perception of the situational, institutional, dispositional, and epistemological barriers to persistence," *Distance Education,* vol. 2, no. 14, pp. 181-198, 1993.

[7] C. Nelms, ". Strengthening America's historically black colleges and universities: A call to action.," in *Paper presented at the Centennial Symposium – Setting the Agenda for Historically Black Colleges and Universities June 2–4, 2010*, Duhram, NC, 2011.

[8] R. Burke and M. Mattis, "Women and minorities in science, technology, engineering, and mathematics:," *Upping the numbers.,* 2007.

[9] N. S. Foundation, "Science and engineering indicators, 2000-2015," 2018. [Online]. Available: http://www.nsf.gov.

[10] R. S. Chen, R. C. Wu, . and J. Y. Chen, " Data mining application in customer relationship management of credit card business.," in *Proceedings of the 29th Annual International Computer Software and Applications Conference, 1-2.*, 2005.

[11] S. M. Metagar, P. D. Hasalkar and S. A. Naik, "Case study of data mining models and warehousing.," *International Journal of Innovative Research in Computer and Communication Engineering, 2(5),*, vol. 2, no. 5, pp. 4408-4412, 2014.

[12] R. Atkinson, J. Hugo, D. Lundgren and M. Shapiro, "Addressing the STEM challenge by expanding specialty math and science high schools.," Information Technology and Innovation Foundation., Washington, DC: , 2007.

[13] H. T. Frierson, J. H. Wyche and W. Pearson, Black American males in higher education: Research, programs, and academe, Bingly: Emerald Group, 2009, pp. 193-208.

[14] D. Chubin, D., G. May and E. Babco, " Diversifying the engineering workforce," *Journal of Engineering Education,* pp. 73-86, January 2005.

[15] J. Passel, G. Livingston and D. Cohn, "Explaining why minority births outnumber white births," 17 May 2017.

[16] ACSFA, "Pathways to success: Integrating learning with life and work to increase national college completion.," ACSFA press., Washington, DC, 2012.

[17] S. Suitts, "Fueling education reform: Historically Black colleges are meeting a national science imperative.," *Cell Biology Education, 2, 205–206. https://doi.org/10.1187/cbe.03-07-0032*, vol. 2, pp. 295-206, 2001.

[18] J. G. Hendrix, "An analysis of student graduation trends in Texas state technical colleges utilizing data mining and other statistical techniques," *Doctoral dissertation).*, 2000.

[19] L. Rokach and O. Maimon, "Feature set decomposition for decision trees.," *Intelligent Data Analysis,*, vol. 9, pp. 131-158, 2005.

[20] L. Rokach and O. Maimon, Data mining with decision trees: Theory and applications. :, Hackensack, NJ: World Scientific Publishing., 2013.