Predictive Analytics Models for Student Admission and Enrollment

Jared Cirelli[1], Andrea M. Konkol[2] and Faisal Aqlan[3]
[1,3]Industrial Engineering Department
[2]Admissions and Financial Aid
Penn State Erie, The Behrend College
Erie, PA 16563, USA
FUA11@psu.edu

Joshua C. Nwokeji
Computer and Information Science Department
Gannon University
Erie, PA, 16541, USA
Nwokeji001@gannon.edu

Abstract

Increasing student admission and enrollment, especially in engineering and computing programs, is a desirable goal for many universities. At the same time, this goal can be difficult to achieve. The aim of this research is to develop a data analytics model that can be used by universities and colleges to improve student admission and enrollment process. Predictive analytics is the technique of using historical data to create, test and validate a model to best describe and predict the probability of an outcome. In recent years, predictive analytics has been used in many areas including manufacturing, healthcare, and service industry. In engineering and computer science education, data analytics models can be used to describe and predict what will happen during the different stages of the enrollment process. This can help an institution determine the interventions that should be taken to support students or meet recruiting goals. In this innovative practice paper, we develop analytics models based on logistic regression, neural networks, and decision trees utilizing historical data from a local university. We focus on the analysis and modeling of student admission and enrollment data to provide a decision support for the admission staff. It may be noted, however, that this model cannot be stand alone and only serves to compliment university administrators' decision-making process to manage admissions and enrollments effectively. The developed models are tested and validated using k-fold cross validation technique.

Keywords
Student admissions; student enrollment; descriptive analytics; predictive analytics data analytics

1. Introduction

Admission and enrollment are among the critical success factors in universities and colleges (hereafter referred to as institutions); since they determine the existence and contribute to the financial viability of many institutions. Moreover some institutions, especially those that are not solely funded by the government or those that don't have other sources of income such as endowments and grants, may want to increase the number of students they admit into their various academic programs while providing adequate resources (Abelt et al., 2015). In addition to increasing the quantity of students, there is also the challenge of admitting quality students, who are able to successful complete their programs and graduate in time. These challenges are even exacerbated in countries like USA, where there are many competing institutions. For instance, between 2010 and 2011, there are 4,599 accredited institutions in the USA (Snyder and Dillow, 2012). These institutions compete to attract a greater number of the best and brightest (quality) students.

Admission challenges are complemented by enrollment challenges which also have budgetary and other implications (Abelt et al., 2015). Normally, not all offers made by institutions' admission office are accepted by students. Students who reject offers may not enroll into any academic program. As noted in Abelt et al. (2015), if lesser than anticipated students enroll, the budget may suffer. Equally, if higher than anticipated enroll, the institutions' facilities such as dorms and learning resources may suffer as well. Therefore, it is equally challenging but necessary to predict the exact number of students that will enroll into various academic programs in an institution.

In order to be viable, institutions must meet their enrollment target to balance their resources and budget; and at the same time, increase the quantity and improve the quality of admitted students to remain competitive Barthelson et al. (2014). Although these conditions for viability are desirable, they are very challenging. There is a plethora of factors that determine which institution a student eventually selects. An institution's accreditation status, recognition of certain specializations, its physical location, campus activities, prominence in sports, etc. are all influencing factors. But these factors, in general, are not controllable and are not considered as attributes of a student. Whereas factors such as performance in high school, test scores, financial aid, race, gender, alumni connections, etc., can be considered as student attributes and hence may turn out to be good predictors of a student's decision to enroll or not.

Traditionally, institutions have advertised themselves by posting information on their websites and using multimedia. However, these traditional methods are increasingly becoming insufficient on their own (Lindbeck and Fodrey, 2010). Therefore, they should be supported by predictive analytics approach that utilizes personal attributes to appeal to the interests of prospective students; and predict the probability that students will accept an offer and enroll into a course. More so predictive analytics can provide accurate information and knowledge about future admission trends, and thus support planning, resource allocation, and decision making regarding the growth of an institution. If the school has an anticipation for a growth in student enrollment then they can plan accordingly to provide adequate resources required to educate students.

This research aims to develop analytics model to predict the probability that student enrollment into a program; and in addition, enhance the quantity and quality of student admission. The data used for this purpose were collected from the admissions office in a local institution. The collected data relate to the following personal attributes of students: intended majors, academic plans, alumni connections, residency, gender, and proximity (i.e., distance from their home to the campus). However, data regarding the students' grade and financial aid was not collected, due to privacy concerns. We then analyzed these data to identify any existing patterns and develop predictive analytics models. Using this models, we predict the anticipated actions, interests, and the chances of students to enroll into a program in our institution. This model can be adopted and customized by institutions to predict the number and quality of admissions and enrollment. The remainder of this paper is organized as follows. In Section 2, we review existing literature related to predictive analytics. This is followed by the description of our framework in Section 3. In Section 4 we discuss the results generated by our approach; and concludes the research with recommendations and future work in Section 5.

2. Related Literature

The general idea of predictive analytics models is to use historical data to predict the future value of an outcome based on one or more input variables. Predictive analytics have been used to assist decision making in different fields including healthcare, manufacturing, service, and academia. Several studies have discussed the development of predictive analytics models for decision making in higher education.

For example, Haddawy developed a predictive model in Haddawy (2007) and used this model to predict the probability of student enrollment. A similar approach described in Abelt et al. (2015) was used to predict the total undergraduate enrollment in University of Virginia. The approaches in Haddawy (2007) and Abelt et al. (2015) appear to make meaningful contributions and may indeed be useful. Only Bayesian Network Algorithm was used in Haddawy (2007), while the approach in Abelt et al. (2015) used three algorithms. In this research, we use six different predictive analytics algorithms and compare them based on their accuracy. The model that has the best accuracy will be recommended.

The predictive model described in [13] and presented at Frontiers in Education Conference (FIE) in 2017, used transcript analysis to support enrollment and predict student success in Computer Science. In a similar approach, Timer and Clauson (2011) used regression and correlation analysis to develop a predictive analytics model. This model was used to predictive student success and admission into nursing programs. The previous two studies presented useful predictive analytics models that can be used by institutions to predict student success. Yet, the data sets used in these studies have a few number of personal attributes and variables. For instance, student GPA is the only attribute considered in Timer and Clauson (2011), likewise the predictive model developed in [13] considers only three attributes namely, student grade, race and gender. The use of limited number of student attributes as variables in the predictive model may produce results that are not optimal, accurate and robust. This is because the algorithms applied in predictive analytics depend on and use attributes as input variables to learn, model and predict future behaviors.

A study reported in Chen et al. (2018)S investigated a college admissions problem in which students' choice set is limited by the strategic actions of the college that are undertaken to prevent multiple applications. The study indicated that when uncertainty is high and the prestige difference is low, a lower-ranked college can obtain more desired students through the strategic actions. An empirical evaluation of the college admission policies on the performance of high school students was discussed (Veloso, 2018). A rank-order tournament model was developed where high school students decide the level of effort and whether or not to take the college admissions test, taking into consideration how these decisions affect their future university admissions chances.

In another study, Baiou and Balinski (2004) developed a graphical modeling of student admissions and faculty recruitment problems. An analysis of the relationship between the timing of college enrollment, educational outcomes, and the timing of family formation decisions in early adulthood was discussed in Humlum et al. (2017). Regression analysis was used to estimate the effect of being above the admission requirements on college enrollment decisions. A study and analysis of admission data and student educational outcomes is presented in Heinesen (2018). A regression discontinuity design was utilized and it was found that being admitted to the first choice program increases the probability of pursuing a master degree in the same subject by 20%. However, the study found no clear evidence that being admitted to one of the higher degree programs listed on the application has an impact on years of education. In Li and Weisman (2011), a simple model was developed that includes three classes of college admissions (i.e., merit, race, and legacy) to study the impact of admission policies on the perforce of students.

In this paper, we propose predictive analytics models to predict the student admission and enrollment. Unlike the existing studies reported above, our approach is based on several input variables mentioned in shown in Table 1. Furthermore, we compare six different predictive analytics models based on their accuracy.

3. Research Framework

The proposed data analytics framework for student admission is shown in Figure 1. The raw data are first analyzed and coded then preprocessed to identify outliers and missing data. Descriptive analytics was used to describe and visualize the data. Feature election was then used to identify the features to be used in the prediction models. The prediction models used include Neural Networks, Logistic Regression, Bayes Net, Random Trees, Chi-square Automatic Interaction Detector (CHAID), and Support Vector Machines. The models were then evaluated based on accuracy, sensitivity, and specificity.

A. Data Collection and Analysis

Data were obtained from the student admission office in a local university. The data were collected for three years (2016 – 2018) and it includes eighteen columns that represent the student and academic programs information. 7345 data points were obtained where each data point represents a student. The data fields along with a brief description of each filed is shown in Table 1. The following Sections will discuss the descriptive and predictive analytics models for the students admission and enrollment data. Figure 2 shows sample graphs for the distributions of the input variables and the types of data measurement scales. The graphs can also be used to data validation. For example, the input variable "adult" should not be considered in the analytics models because the majority of the data points belong to one category and this will impact the accuracy of the results. Hence, this variable will be excluded.

B. Descriptive Analytics

Descriptive analytics is a preliminary stage of data processing. The purpose of descriptive analytics is to create a summary based on data or a dataset. The summary often provides useful insight such as patterns and trends that analysts will be able to use to predict the occurrences of an event. Figure 3 shows the distribution of the students who applied to the university based on their gender and underrepresented status. IBM SPSS Modeler® software was used to perform the analysis of the data. IBM SPSS Modeler® is a data mining and text analytics software that is used to build predictive and descriptive analytics models.

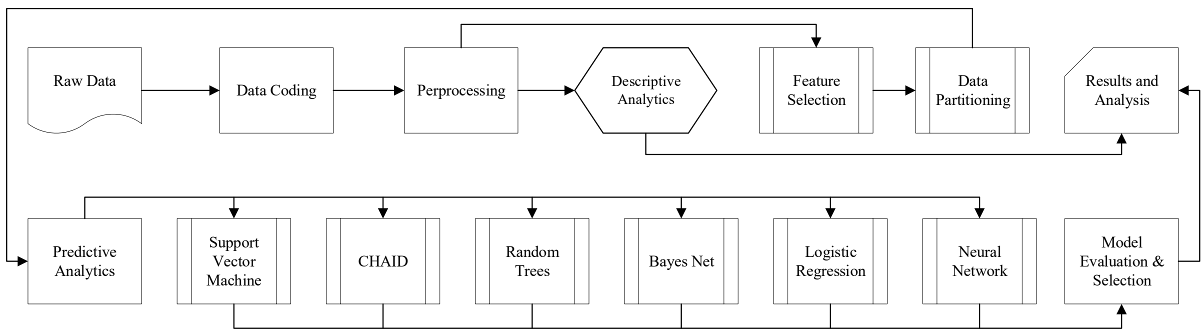

Figure. 1. Data analytics framework

Table 1. Student Admission data

Parameter	Explanation
Acad Program	Career and college code
Acad Plan	Student's current plan, college and major status
Intended Plan	Applicant's intended major
Status/Action: CN/WADM	Offer expired, admitted and did not attend
Status/Action: CN/WAPP	Applicant notified school they will not attend
Status/Action: AC/MATR	Applicant is at the school as a student
Status/Action: PM/DEIN	Admitted and paid acceptance fees but did not matriculate, likely not at school this semester
Admit Term Descr	Semester term that applicant is admitted
First Choice Campus	Applicant's first desired campus
Alt Choice Campus	Applicant's alternative desired campus
Underrepresented	Minority, any race other than white
Alumni Connection	Family member that attended univ.
Adult	Applicant over age of 26
Residency	In State or Out of State
HS City	City where applicant attended High School
HS State	State where applicant attended High School
Region	Current region of applicant
Appl Ctr	Domestic or International

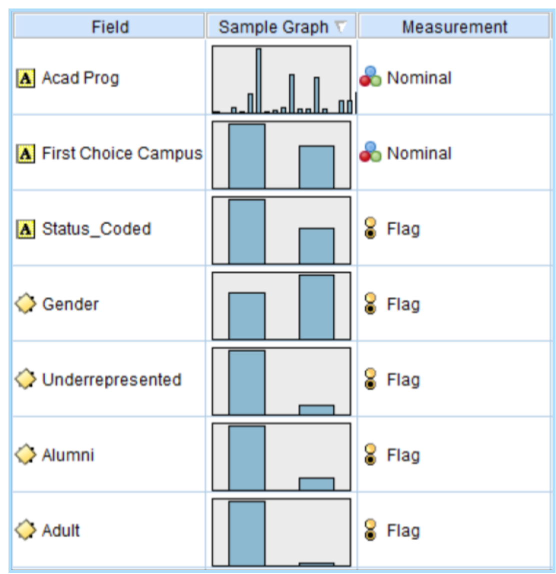

Figure 2. Distributions and types of mesurement for the input variables

Figure 4 shows the number of students per the enrollment status where "AC" represents the students who were admitted and currently enrolled and "CN" represents the students who were admitted but are not currently enrolled. As indicated in the figure, about 34% of the admitted are enrolled whereas 66% of the admitted students and not currently enrolled. These students may have moved to other composes or other universities. Figure 5 shows the distribution of the students based on residency (i.e., in state and out of state students). "INST" stands for in state students and "OUST" stands for out of state students. There are some missing data points and these data points were removed to improve the accuracy of prediction.

C. Predictive Analytics

Predictive analytics models are developed to predict the student enrolment status based on input factors. Several predictive analytics methods were utilized including: Artificial Neural Networks (ANN), Logistic Regression (LR), Bayesian Networks (BN), Random Trees (RT), Chi-square Automatic Interaction Detector (CHAID), and Support Vector Machines (SVM), See Table 2 below.

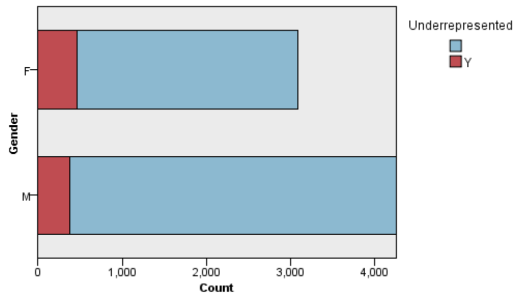

Figure 3. Number of students per gender and underrepresented

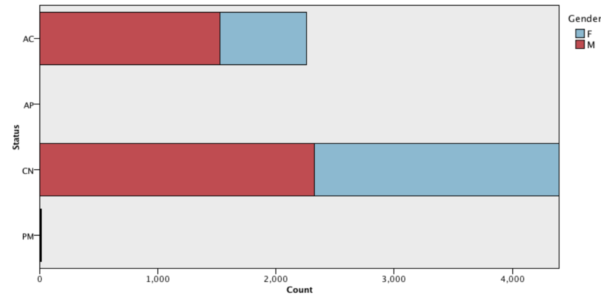

Figure. 4. Number of students per enrollment status

ANN is a simplified model of the way the human brain processes information. It works by simulating a large number of interconnected processing units that resemble abstract versions of neurons. The processing units are arranged in three layers: input layers, hidden layers, and output layers.

LR is a statistical technique that classifies the output based on the values of the inputs. LR is a form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictor. The general form the LR equation is shown below:

$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

where P is the probability of the output, β is the coefficient of the output variables, and x is the output variable.

BN is a graphical model that shows the variables of a dataset and the probabilistic independencies between them.

RT is a classification (and prediction) method that is built on Classification and Regression Tree methodology. It uses recursive partitioning to split the training records into segments with similar output field values.

CHAID is a classification method for building decision trees by using chi-square statistics to identify optimal splits. It is used to discover the relationship between variables by building a predictive model to determine how variables best merge to explain the outcome in the given output variable.

SVM is a supervised learning models that uses learning algorithm to analyze data based on mapping data to a high-dimensional feature space so that data points can be categorized. SVM is usually used with data that have a large number of predictor variables.

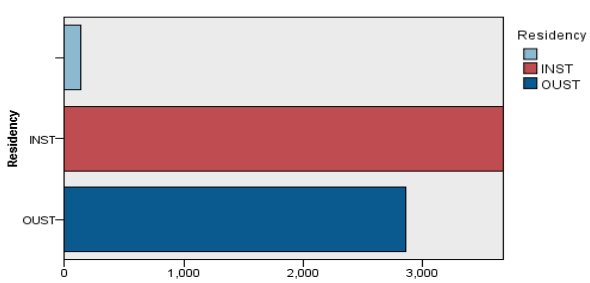

Figure 5. In state and out of state students

Table 2. Predictive analytics methods

Model	Description
ANN	Mimics human brain functions to identify underlying relationships
LR	Independent variables are used to determine an outcome.
BN	Set of variables and their conditional dependencies via a directed acrylic graph
RT	Explains how variables best merge to explain an outcome
CHAID	Discriminative classifier that categorizes data
SVM	Machine learning classifier based on random subsets of variables for each tree. The most frequent tree output is used as overall classification

Table 3 shows the confusion matrix or the classification table which consists of four outcomes True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). The outcomes are used to estimate the performance measures of the models.

Table 3. Confusion matrix

Test Outcome		Predicted	
		Positive (1)	Negative (0)
Observed	Positive (1)	TP	FP
	Negative (0)	FN	TN

The predictive analytics models are evaluated based on three performance measures: accuracy, sensitivity, and Specificity. These performance measures are used to compare the model and identify the best model for student admission and enrollment prediction. The definitions of the three performance measures is shown in Table 4.

The data was partitioned into 80% (i.e., 5316 data points) for training and 20% (i.e., 1333 data points) for testing. The results of the predictive models are shown in Tables 5 and 6. For both training and testing data, the accuracy of the prediction models was calculated. The average prediction performance of the models is about 72%. Models performance can be improved by considering other input variables such as testing scores, high school grades, and financial aid status. Such variables were not available in our data because the university considers this as confidential information. Moreover, more data can be collected and used to enhance the production accuracy.

Table 4. Performance metrics for evaluating the analytics models

Metric	Description	Equation
Accuracy	Measures the ability of the model to correctly predict the class label of new or unseen data.	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	Measures the proportion of positives (or Yes's) that are correctly identified as such.	$\frac{TP}{TP+FN}$
Specificity	Measures the proportion of negatives (or No's) that are correctly identified as such.	$\frac{TN}{TN+FP}$
Abbreviation	Name	Description
TP	True Positives	Number of correct classifications predicted as positive (or Yes)
TN	True Negatives	Number of correct classifications predicted as negative (or No)
FP	False Positive	Number of examples that are incorrectly predicted as positive when it is actually negative
FN	False Negative	Number of examples that are incorrectly predicted as negative when it is actually positive

Table 5. Performance measures for the training data

Method	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Neural Network	1036	2806	790	684	72.3%	60.2%	78.0%
Logistic Regression	815	3031	1011	459	72.3%	64.0%	75.0%
Bayes Net	1064	2837	762	653	73.4%	62.0%	78.8%
Random Trees	1318	2328	1096	486	70.0%	73.1%	68%
CHAID	1070	2820	756	670	73.2%	61.5%	78.9%
Support Vector Machines	828	2979	998	511	74.0%	61.8%	74.9%

Table 6. Performance measures for the test data

Method	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Neural Network	229	778	203	123	75.4%	65.0%	79.3%
Logistic Regression	230	787	202	117	76.1%	66.3%	79.6%
Bayes Net	225	746	177	158	74.9%	58.8%	80.1%
Random Trees	291	624	133	266	70.0%	52.2%	82.4%
CHAID	210	784	222	120	74.6%	63.6%	77.9%
Support Vector Machines	223	803	209	101	76.8%	68.8%	79.3%

Figure 6 shows the structure of the probabilistic graphical model of the Bayesian Network. The output variable is the "Status_Coded" which represent the student admission and enrollment status. The output variable is linked to the eight input variables. The figure shows the conditional dependencies via the directed acyclic graph.

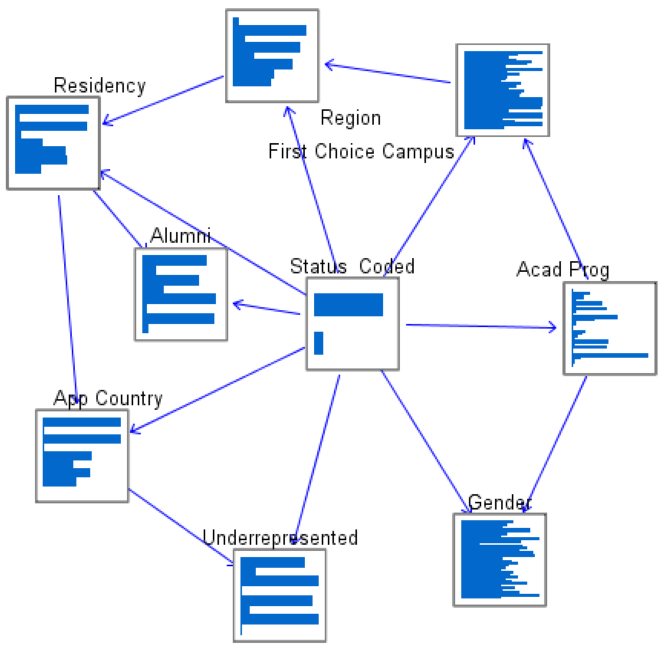

Figure 6. Baysian Network model

Figure 7 shows the decision tree generated by the CHAID method. The top node represents the output variable, "Status_Coded", which has two categories for the student admission and enrollment staus. Figure 8 shows the structue of the Neural Network model. The figure shown the eight input variables, neurons, and the ourput variable.

The descriptive and predictive analytics models developed in this study can be used by decision makers for planning purposes. Specifically, the predictive analytics models can be used to predict the number of students who will join the university given their application information. The accuracy of the predictive models needs to be imporved by collecting more historical data.

4. Conclusions

This paper discussed the development of analytics models for student admission and enrollment in a local university. Eight input variables were used to predict the student enrollment using different analytics models. The overall average accuracy of the models is 72%. The results can be used by the student admission and enrolment office for planning purposes. The outcomes can be used to improve student retention.
Future work will focus on improving the accuracy of the models. Further information such as testing scores, high school grades, and financial aid status could be used to improve accuracy of the models.

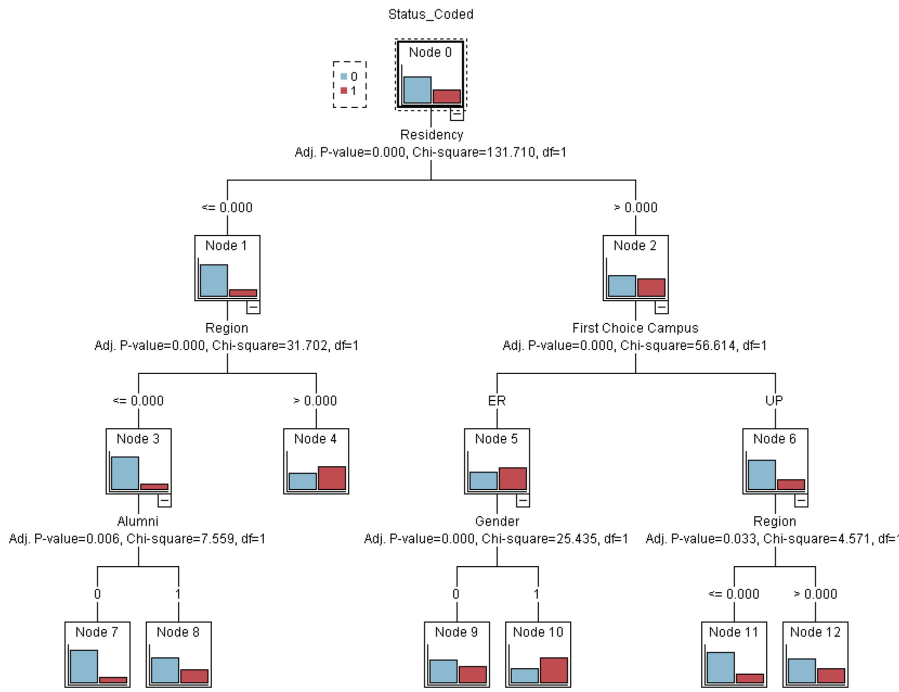

Figure 7. Decision Tree

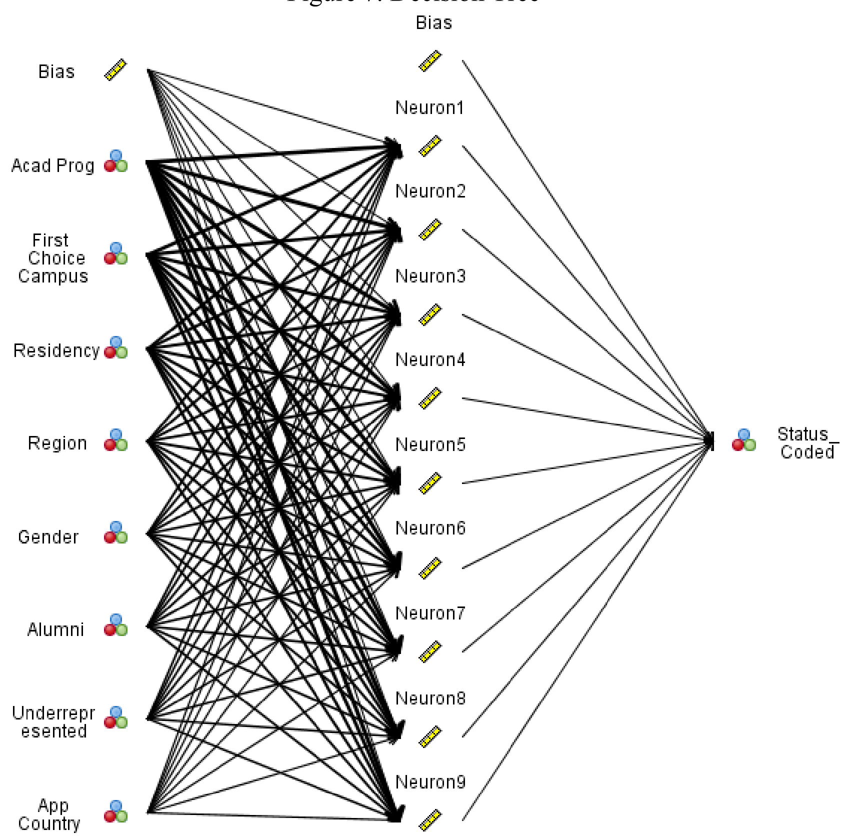

Figure 8. Neural Network model

References

Timer, J.E., and Clauson, M.I., The use of selective admission tools to predict students' success in an advanced standing baccalaureate nursing program, *Nurse Education Today*, vol. 31, pp. 601-606, 2011.

Chen, W.C., Chen, Y.Y., and Kao, Y.C., Limited choice in college admissions: An experimental study, *Games and Economic Behavior,* vol. 108, March, pp. 295-316, 2018.

Veloso, N.G., The impact of college admissions policies on the academic effort of high school students, *Economics of Education Review*, vol. 65, August, pp. 58-92, 2018.

Baiou, M., and Balinski, M., Student admissions and faculty recruitment, *Theoretical Computer Science*, vol. 322, no. 2, pp. 245-265, 2004.

Humlum, M.K., Kristoffersen, J.H.G., and Vejlin, R., College admission decisions, educational outcomes, and family formation, *Labour Economics*, vol. 48, October, pp. 215-230, 2017.

Heinesen, E., Admission to higher education programmes and student educational outcomes and earnings-evidence from Denmark, *Economics of Education Review*, vol. 63, April, pp. 1-19, 2018.

Li, D., and Weisman, D.L., Why preferences in college admissions may yield a more-able student body, *Economics of Education Review*, vo. 30, no. 4, pp. 724-728, 2011.

Snyder, T.D., and Dillow, S.A., Digest of Education Statistics, National Center for Education Statistics, 2012.

Abelt, J., Browning, D., Dyer, C., Haines, M., Ross, J., Still, P., a Gerber, M., Predicting likelihood of enrollment among applicants to the UVa undergraduate program, *Systems and Information Engineering Design Symposium*, pp. 194-199, 2015.

Barthelson, M., Boumlic, I., and Shamma, U., Design to improve the freshman admissions process, *Systems and Information Engineering Design Symposium*, pp. 124-128, 2014.

Lindbeck, R., and Fodrey, B., Using technology in undergraduate admission: A student perspective, *Journal of College Admission*, vol. 208, pp. 10-17, 2010.

Haddawy, P., Deriving financial aid optimization models from admissions data, *Frontiers in Education Conference-Global Engineering: Knowledge without Borders, Opportunities without Passports*, 2007.

Trytten, D. A., and McGovern, A., Moving from managing enrollment to predicting student success, *Frontiers in Education Conference (FIE)*, pp. 1-9, 2017.

Biographies

Jared Cirelli is an undergraduate students in the Industrial Engineering department at Penn State Erie, The Behrend College.

Andrea M. Konkol is the Associate Director of Admissions at Penn State Erie, The Behrend College. She earned her M.B.A. from Penn State Erie, The Behrend College in 2003.

Faisal Aqlan is currently an assistant professor of Industrial Engineering and Master of Manufacturing Management (MMM) at Penn State Behrend. He earned his Ph.D. in Industrial and Systems Engineering from the State University of New York at Binghamton in 2013. Aqlan has worked on industry projects with Innovation Associates Company and IBM Corporation. His work has resulted in both business value and intellectual property. He is a certified Lean Silver and Six Sigma Black Belt. He is a senior member of the Institute of Industrial and Systems Engineers (IISE) and currently serves as the president of IISE Logistics and Supply Chain Division, director of Young Professionals Group, and founding director of Modeling and Simulation Division. Aqlan is also a member of American Society for Quality (ASQ), Society of Manufacturing Engineers (SME), and Industrial Engineering and Operations Management (IEOM) Society. He has received numerous awards including the IBM Vice President award for innovation excellence, Penn State Behrend's School of Engineering Distinguished Award for Excellence in Research, and the Penn State Behrend's Council of Fellows Faculty Research Award. Aqlan is the Principal Investigator and Director of the NSF RET Site in Manufacturing Simulation and Automation at Penn State Behrend.

Joshua C. Nwokeji is an emerging professor, researcher and entrepreneur. He completed his Ph.D., in June 2016, and joined Gannon University, as Assistant Professor of Information Systems, in August 2016. He has authored more than 13 peer reviewed publications, and served as a reviewer in international conferences such as HICSS, ICEIS, and AMCIS. Currently he is a member of editorial board of ORBIT Journal. Dr. Nwokeji has taught many courses including systems analysis and design, database management systems, requirements engineering, and business process modeling.