Employing Various Data Mining Techniques to Forecast the Success Rate of Information Technology Education Students

Mosbah Mohamed Elssaedi

mosbah_us@hotmail.com

Department of Computer Science, Faculty of Science, Sirte University - Libya

ABSTRACT

This study was designed to investigate the factors that affect the success rate of Information Technology Education students which composed of Computer Science and Information Technology. Several variables such as years of graduation, entrance exams, and other variables have been used for the investigation. Several data mining techniques such as linear regression, neural network, and decision tree have employed to determine the valid predictors and acceptability of the data mining technique. The results show that the best predictor taken from the entrance exams is non-verbal ability while the best forecasting using data mining is decision tree analysis with 99.19 percent accuracy. If the results taken from the system will be incorporated in entrance examinations results, admission office will be able to identify students that can graduate on-time and whose students should be taken as probationary in the program. It can also identify students not to be taken in the program to avoid waste of time in studying at the University.

Keyword— Neural network, linear regression, decision tree, forecasting, data mining.

1 Introduction

Technology is the application of scientific knowledge for practical purposes especially in industry, engineering and applied sciences. Technology can be used at work to extract materials, transportation, learning, manufacturing, creating artefacts, securing data, scaling business and more [1]. The benefits of technology in education have been a gateway to the new learning environment. Assistive technology that helps children with special needs such as an e-reader, adaptive voice software are a few good examples. The use of computers and innovation in classrooms has opened up an entirely new system for showing and successful learning. Computers generate huge amounts of information and can benefit the field of education, like helping students learn faster or make learning more interesting to some extent.

With technology's rapid development, no surprise that there is an abundant of courses related to technology that is offered by different university worldwide.

The major problem is that at there is a big percentage of the unemployed with an average of 17% of college graduate came from Bachelor of Science in Computer Science and Bachelor of Science in Information Technology [2]. This sort of measurement is disturbing; it implies that colleges are delivering graduates with lacking abilities with the connection to Computer Science and IT which add to the unemployment rate. National development is every nation's goal throughout the world. A country is seen to be developed when underemployment and unemployment rates decrease if not eliminated. One of the probable reasons for this is the mismatch between education and employment.

Admission Examination or University Entrance Examination is believed to help students to select an appropriate course in college matching their aptitude and maintains a quality of education thereby, bringing national development in the country. Among the objectives of the creation of admission, examinations are to minimize aimless wastage of labour and different assets which generally could be coordinated towards more beneficial endeavours. It also assesses the capabilities and skills the students develop in their early studies which are necessary to be successful in college or even becoming an entrepreneur [3]. It can be very beneficial for students if followed but many students choose to deviate from the recommendations made by Admission by choosing another program in college. The huge amount of information and data, when analyzed can help in decision making and help to create a model that determines the success rate of Computer Science and IT student deviating their admission exams results using data mining. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events and outputs useful information while reducing the quantity of data [4]. This study aims to develop a model that determines the success rate of computer science and IT student that deviated from a recommendation from Admission result. Furthermore, to find out the field of interest that should be improved to pass computer science or IT degree and help the academe reduce drop-outs and shifters of the program offered.

This study was designed to investigate the factors that affect the success rate of Computer Science and IT student that deviates from the recommendation of admission examination. It will also correlate to the abilities of students in the result of admission examinations to the date of graduation of Computer Science and IT students and to develop a model that determines the success rate of students. Specifically the research will be guided by the following questions: what are the predictors to be considered to determine the success rate of Computer Science and IT student that deviate from the recommendation of admission examinations?, what are the correlations of the predictors in passing On – Time the Computer Science degree? and what are the data mining techniques and algorithms will be used in the forecasting model?

2 Related Literature

Career assessment is the first step in career planning. Admission Test or College Entrance Examination is one of the career assessments given by the any University to entrants students. There is a study that aims student's performance base on information's like attendance, class test, name, age, course, and topographic location or address. In this research, they used the classification method of data mining using a decision tree. A forecasting system model was developed in a study in which the main goal is to analyze the learning behaviour of the higher education students[5]. They used the decision tree, a popular classification method of data mining that results in a flowchart-like a tree structure where each node denotes a test on an attribute value. Another forecasting system model was developed which the main goal is to understand student data such as name, age, address, student's grades and course for career selection and job absorption rate after graduation[6]. There is research that finds out whether natural talents and interest of 116 students based on admission exams result match the program that the students have enrolled in. In this study, the researcher employed a qualitative research design. There are eight potentials/inclinations of students measured: scientific ability, reading comprehension, verbal ability, mathematical ability, clerical ability, manipulative skills, nonverbal ability and entrepreneurial skills[7]. In the conducted study, the results show that the respondents have varied occupational interest based on an admission exam. Based on this research, most of the respondents choose not to enrol in the program that matches respondent's field of interest that leads to the respondent's failure on some of the respondent's subject while the respondents that enrol in the program that matches the respondent's field of interest were seen to have become successful.

3 Methods

In building the forecasting model, student data such as admission examination field of interest, occupational Interests and the student's date of graduation were used to find the relationship by feeding to the data mining analyzing tool. The data sets that were used to develop the forecasting model are the records of these students of Computer Science and IT degree. The data of Computer Science and IT student batch 2011–2015 served as the training data of the forecasting model and the Computer Science student batch 2016 served as the test data that evaluate the model The data used were combined data from the admission office of one University in the Philippines and compared to the data of Sirte University. The data was laballed according to the data to conform with Sirte University. Rapidminer, a data mining tool is used to process data and create a forecasting model. Certain algorithms are used in creating the forecasting model and to find out the best ability in admission exams that best affect the performance of the Computer Science and IT students.

Descriptive Correlation. The correlation measures the strength and direction of a linear relationship between two variables. The value correlation is always between +1 and -1. The following are an interpretation of values in correlation:

- Exactly -1. A perfect downhill linear relationship.
- – 0.70. A strong downhill linear relationship.
- – 0.50. A moderate downhill linear relationship.
- – 0.30. A weak downhill linear relationship.
- 0. No linear relationship.
- + 0.30. A weak uphill linear relationship.
- + 0.50. A moderate uphill linear relationship
- + 0.70. A strong uphill linear relationship.
- Exactly +1. A perfect uphill linear relationship.

3.1 Conceptual Framework

A conceptual framework was used by the researchers to outline the courses of action or to present a preferred approach to the developed system. As shown in Figure 1, there are five phases in the proposed forecasting model:

1. *Data Gathering* – this is the first phase of the model where data such as the admission examination result were extracted which contains the different fields of interest, the occupational interest of the student and the student graduation date. This data will be saved in the repository.
2. *Data Pre-Processing* – the second phase of the model this is where the data cleansing, estimating of missing values and normalization of database takes place.
3. *Modelling* – the third phase of the model is where the building of the model takes place. The data mining tool will analyze all the data and outputs all data mining algorithms results for each technique.
4. *Determine the Success Rate* – the fourth phase of the model is the entry point of the test data to evaluate the model's accuracy.
5. *Result* – the last phase of the model is where the output will be displayed in a dashboard which is the success rate of computer science and IT students that deviates from the admission results.

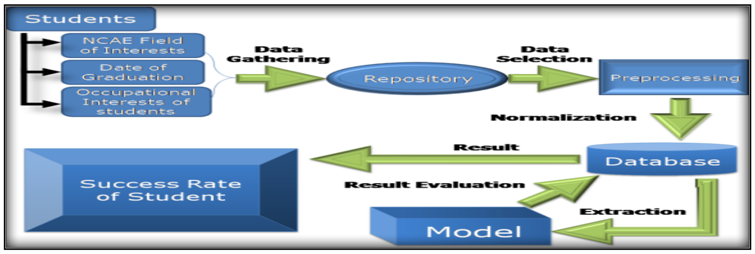


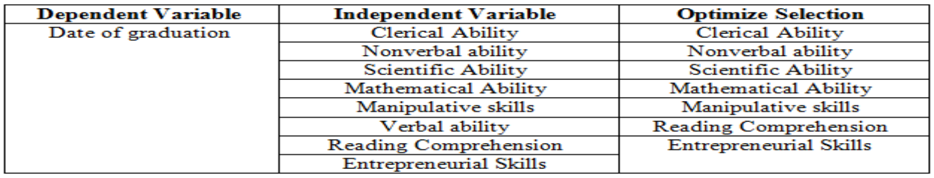
Figure 1: *Phases of the model*

4 Results and Analysis

4.1. Predictors for Success Rate

The variables used in the study are the data of Computer Science and IT students from the school year 2011– 2015 with 368. The data set is put in the outliner detection to minimize the noise of the actual datasets and to increase the accuracy of the forecasting model before measuring the significance of each predictor.

Table 1: *Predictors and variables*



Dependent Variable	Independent Variable	Optimize Selection
Date of graduation	Clerical Ability	Clerical Ability
	Nonverbal ability	Nonverbal ability
	Scientific Ability	Scientific Ability
	Mathematical Ability	Mathematical Ability
	Manipulative skills	Manipulative skills
	Verbal ability	Reading Comprehension
	Reading Comprehension	Entrepreneurial Skills
	Entrepreneurial Skills	

Table 1 shows the predictors and the variables used in the study. The researcher used the date of graduation of the students as the dependent variable because the student's date of graduation defines if the student is successful in the course that the students take during college. After the detection of outliners, the researchers used the optimize selection operator in RapidMiner. It selects the most relevant attributes of the given datasets. Two deterministic greedy feature selection algorithms (forward selection and backward elimination) are used for feature selection. It eliminated the Verbal Ability using backward elimination from the seven dependent variables.

4.2 Correlation of Predictors to Graduate On-Time in Computer Science Degree

The researchers used the correlation of RapidMiner to measure the relationship and strengths of the variables after feeding and using correlation in the tool. Table2shows the correlation between On-Time graduations of a student with Abilities extracted from admission exams. The scientific ability has a moderate uphill linear relationship with On-Time graduations of Computer Science and IT student with 0.430, meaning it has a moderate effect on the success rate of passing the degree. Table 3 shows that the most important ability in admission exam to pass Computer Science and IT program is the Non Verbal Ability.

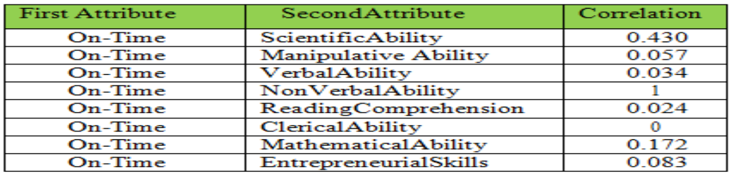
4.3 Data Mining Techniques and Algorithms

The researcher used three data mining techniques and compared each result to find out what are the best data mining techniques that best suits the data sets that the researchers gathered namely regression, neural network, and decision tree analysis.

4.3.1 Regression

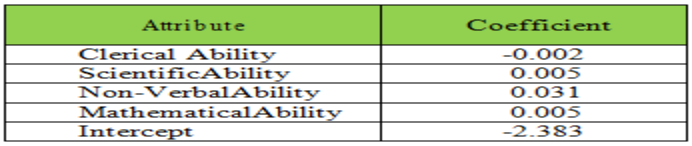
The researcher used linear regression to be able to forecast the success rate of computer science student in passing the computer science degree using the abilities presenting NCAE. After applying the model to the RapidMiner tool, the researcher came up and produced the following output:

Table 2: *Correlation predictors*



First Attribute	SecondAttribute	Correlation
On-Time	ScientificAbility	0.430
On-Time	Manipulative Ability	0.057
On-Time	VerbalAbility	0.034
On-Time	NonVerbalAbility	1
On-Time	ReadingComprehension	0.024
On-Time	ClericalAbility	0
On-Time	MathematicalAbility	0.172
On-Time	EntrepreneurialSkills	0.083

Table 3: *Linear regression output*



Attribute	Coefficient
Clerical Ability	-0.002
ScientificAbility	0.005
Non-VerbalAbility	0.031
MathematicalAbility	0.005
Intercept	-2.383

Table 3 was the result after using the linear regression from the RapidMiner tool, the coefficients are used for Linear Regression Model which is:

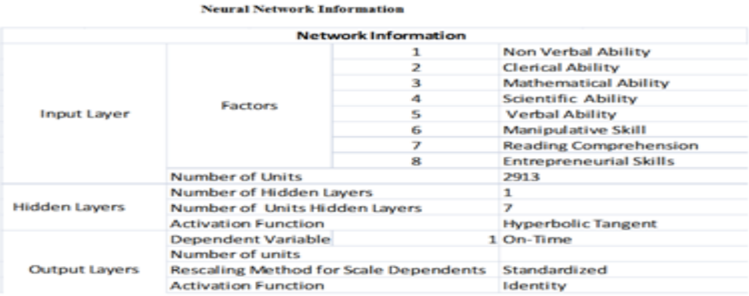
$$Y = -2.383 + (-0.002)\ (\text{Clerical Ability}) + (0.005)\ (\text{Scientific Ability}) + (0.031)\ (\text{Non Verbal Ability}) + (0.031)\ (\text{Non Verbal Ability}) + (0.005)\ (\text{Mathematical Ability})$$

To get the forecasted value, the intercept should be subtracted to the sum of products of abilities in admission exams to their respective coefficient values. If the value is equal to 1, the student will pass, and if it is 0, the student will not pass the CS degree.

4.3.2 Neutral Network

The researcher used a neural network for the next test and Table 4 shows the three layers under Multilayer Perception of the neural network. In able to map the following input, The researcher uses the predictors as the input layer or factors consists of Non-Verbal Ability, Clerical Ability, Mathematical Ability, Scientific Ability, Manipulative Skills, Verbal Ability, Reading Comprehension and Entrepreneurial skill. The hidden layer consists of Number of units, Number of Hidden Layers, Number of Units in the hidden layer, Activation Function, Dependent Variables and Number of Units. The last year or the output layer consists of Rescaling Method for Scale Dependents, Activation Function, and Error Function. The neural network normalized the data and was used to map and get the input layer, hidden layer, and output layer. It was developed to measure the frequency of action potentials.

Table 4: *Neural network analysis*



Neural Network Information

Network Information			
Input Layer	Factors	1	Non Verbal Ability
		2	Clerical Ability
		3	Mathematical Ability
		4	Scientific Ability
		5	Verbal Ability
		6	Manipulative Skill
		7	Reading Comprehension
		8	Entrepreneurial Skills
	Number of Units		2913
Hidden Layers	Number of Hidden Layers		1
	Number of Units Hidden Layers		7
	Activation Function		Hyperbolic Tangent
Output Layers	Dependent Variable	1	On-Time
	Number of units		
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity

4.3.3. Decision Tree Algorithm

Figure 2 shows the result of the decision tree using CHAID algorithm with the predictors are the child and the dependent variable was the parent. The researchers used a classification tree to classify the nodes of the predictors according to the data. The Node 0 which is Non Verbal ability is the highest factor or the best predictor to determine if the student will pass the Computer Science or IT degree. The node of the Non-Verbal Ability was the computed quantitative measurement of the data of graduates of computer science and IT program in the University.

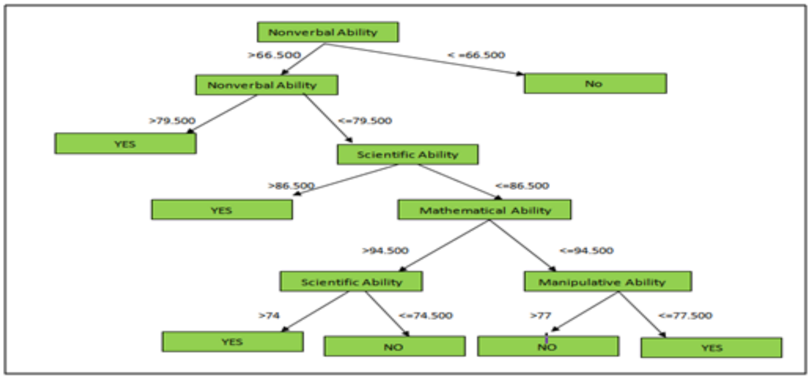
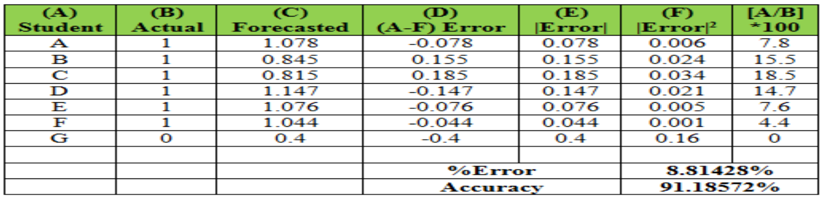


Figure 2: *Decision tree analysis*

4.4 Level of Acceptability of the Forecasting Model.

Accuracy or decision effectiveness is the main factors in evaluating a forecasting model[8]. The RapidMiner tool has a performance evaluator who can display the accuracy, precision, and area of the curve in each model created using the tool.

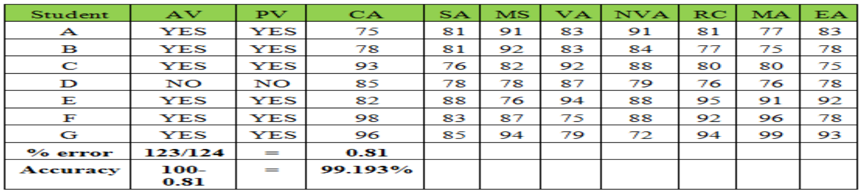
Table 5: S*ample measure of forecasted result: linear regression*



(A)	(B)	(C)	(D)	(E)	(F)	[A/B]
Student	**Actual**	**Forecasted**	**(A-F) Error**	**|Error|**	**|Error|^{2}**	***100**
A	1	1.078	-0.078	0.078	0.006	7.8
B	1	0.845	0.155	0.155	0.024	15.5
C	1	0.815	0.185	0.185	0.034	18.5
D	1	1.147	-0.147	0.147	0.021	14.7
E	1	1.076	-0.076	0.076	0.005	7.6
F	1	1.044	-0.044	0.044	0.001	4.4
G	0	0.4	-0.4	0.4	0.16	0
			%Error		**8.81428%**	
			Accuracy		**91.18572%**	

Table 5 shows the actual and forecasted value in the batch 2011 of Computer Science and IT student of the University, the percentage of error of student's actual value and the accuracy of the predicted value. The research used different measures such as MAD (|Error|) which weights all errors evenly, MSE (|Error|^2) which weights errors according to their squared values, and MAPE ([|Error|/Actual]*100) which weights according to relative error. The percent error computed is 8.81428%, almost equal to 91.18572%, which means that the result for the model is accurate and therefore the admission exams occupational interest was relative in the success rate of the students.

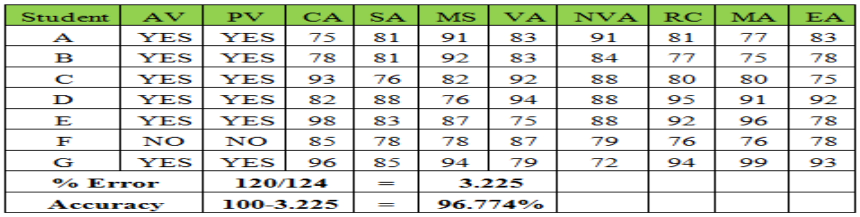
Table 6: *Sample measure of forecasted result: decision tree*



Student	AV	PV	CA	SA	MS	VA	NVA	RC	MA	EA
A	YES	YES	75	81	91	83	91	81	77	83
B	YES	YES	78	81	92	83	84	77	75	78
C	YES	YES	93	76	82	92	88	80	80	75
D	NO	NO	85	78	78	87	79	76	76	78
E	YES	YES	82	88	76	94	88	95	91	92
F	YES	YES	98	83	87	75	88	92	96	78
G	YES	YES	96	85	94	79	72	94	99	93
% error	**123/124**	=	**0.81**							
Accuracy	**100-0.81**	=	**99.193%**							

From the given table, the accuracy of decision tree mode is the difference of the percent error which is 0.81 and 100 percent as shown in Table 6. After solving the level of acceptability, the result is 99.193%, which means that the result for the decision tree was accurate and can be used as evidence of relevance to the study.

Table 7: *Sample measure of forecasted result: decision tree*



Student	AV	PV	CA	SA	MS	VA	NVA	RC	MA	EA
A	YES	YES	75	81	91	83	91	81	77	83
B	YES	YES	78	81	92	83	84	77	75	78
C	YES	YES	93	76	82	92	88	80	80	75
D	YES	YES	82	88	76	94	88	95	91	92
E	YES	YES	98	83	87	75	88	92	96	78
F	NO	NO	85	78	78	87	79	76	76	78
G	YES	YES	96	85	94	79	72	94	99	93
% Error		**120/124**		=	**3.225**					
Accuracy		**100-3.225**		=	**96.774%**					

The accuracy of the neural network model is the difference of the percent error which is 3.225 and 100 percent. After solving the level of acceptability, the result is 96.774%, which means that the success rate of the students is valid with the help of admission exams occupational interest. Based on the result and observations from the model together with data that was being gathered, here are the findings of the researcher.

1. Predictors that are considered to determine the success rate of Computer Science student from the recommendation of Admission Exams. Based on the gathered data the predictors that are considered in predicting the success rate of Computer Science students are Non-Verbal Ability, Verbal Ability, Clerical Ability, Reading Comprehension, Manipulative Skills, Mathematical Ability, Scientific Ability and Entrepreneurial Skills. The abilities that are present in NCAE results are used as a dependent variable or the input to forecast the outcome or independent variable. The researcher used the date of graduation of Computer Science students as the independent variable because this will determine if the student is successful in the Computer Science degree if the student was able to finish the degree within four years, then the student is said to be successful.

2. Relationship of the predictors in passing Computer Science degree. The researcher used the RapidMiner tool to be able to find the value of correlation of the predictors to the success rate of Computer Science student. After feeding the data in the tool, the researcher found out that the Non-Verbal Ability is the best predictor to determine the success rate because it has a perfect linear relationship with the success rate of Computer Science student in passing the degree. Followed by the Scientific Ability which has a moderate uphill linear relationship with the success rate.

3. Data mining techniques and algorithms that were used in the forecasting model. The data mining techniques that were used by the researcher during the study are Linear Regression, Neutral Network, and Decision Tree. Linear regression estimates the value of the target as a function of the predictors for each case. The neutral network which consists of an interconnected group of artificial neurons, and it process information using a connectionist approach to computation. Decision tree shows how one choice leads to the next, and the use of branches shows that every alternative or option is mutually exclusive. The researcher used the different measures such as MAD(|Error|)which weights all errors evenly, MSE(|Error|) which weights errors according to their squared values, and Physical Education[|Error|/Actual]*100 which weights according to relative error. The researcher used Physical Education or Linear Regression; the researchers found out that physical education is the most popular aggregate measure of forecasting accuracy. The percent error computed is 8.81428% or equal to 91.18572% accurate. For the remaining models which are the neural network and decision tree, the researcher used the performance evaluator of Rapid miter. The neural network got an accuracy of 96.774% while the decision tree got an accuracy of 99.193%.

5 Conclusion and Recommendation

1. The proposed study could be a great help to the Admission Office or Registrar of the University students in enhancing their basic forecasting skills especially using Data Mining.

2. The researcher considered that the forecasting model that was made efficient if it will be used to determine the success rate of Computer Science and IT student in passing the degree.
3. There is a significant relationship between the date of graduation of Computer Science and IT student and predictors in forecasting the success rate of the students in passing the Computer Science degree.

The researcher recommends that there should be a separate section for students that have a low score on Non-verbal ability to give more attention or supervision in teaching them because these students have a high chance in failing the Computer Science degree.

References

[1] Fildes, R. and Kourentezes, N. (2010). Validation and Forecasting Accuracy in Models of Climate Change.
[2] Ramey. (2012). Use of Technology.
[3] Mallorca, R. (2008). Student's Natural Aptitudes and the Required Skills in their Chosen Program."
[4] Gupta. S., Adhay Bnasal, and Retish Rastogi. (2012). Learning Behaviour of Analysis of Higher Studies Using Data Mining.
[5] Saurab Pal. (2012). "Mining Educational Data to Reduce Drop-out Rates of Engineering Students.
[6] Torres, T. (2015). I8 percent of unemployed college graduates – NSO.
[7] TuffreyStephane (2011). Data Mining and Statistics for Decision Making Statistic for Dummies.
[8] Wilma L. Labrador. (2009). National Career Assessment Examination (NCAE) As It Influences The TV System.