A Decision Support Tool for Inferring Further Education Desires of Youth in Sri Lanka

M.F.M Firdhous*
Department of Information Technology
Faculty of Information Technology
University of Moratuwa
Moratuwa
Sri Lanka

Ravindi Jayasundara
Department of Mathematics
Faculty of Engineering
University of Moratuwa
Moratuwa
Sri Lanka

*corresponding author

Abstract – This paper presents the results of a study carried out to identify the factors that influence the further education desires of Sri Lankan youth. Statistical modeling has been initially used to infer the desires of the youth and then a decision support tool has been developed based on the statistical model developed. In order to carry out the analysis and the development of the model, data collected as part of the National Youth Survey has been used. The accuracy of the model and the decision support tool has been tested by using a random data sets and the accuracy was found to be well above 80 percent, which is sufficient for any policy related decision making.

Keywords – Educational Desires of Youth; Univariate Analysis; Logit Model; Data Mining.

I. INTRODUCTION

Sri Lanka has witnessed several incidents of youth unrest in the recent past. Out of these insurgencies, two insurgencies involved the youth of the south while the other one involved the youth of the north. There have been many discussions and debates about youth unrest and the increasingly violent and intolerant nature of their conflicts. Since these discussions have been rather impressionistic there has always been the need for systematic studies to obtain information on Sri Lankan youth and their background and desires [1]. In order to collect up to date information, targeting to explore facts on Sri Lankan youth and their perceptions, an island wide national youth survey has been carried out. This has been conducted as a joint undertaking involving the United Nations Development programme (UNDP) and other six Sri Lankan and German institutions in the turn of the century. In this survey they have considered four main segments of youth, that is, their politics, conflicts, employment and education. Further Education Desires of youth have been selected to be studied further in this research. The relationship between the types of further education desire of youth in Sri Lanka had been studied with relation to other social factors.

Education domain consists of many different areas but presently in Sri Lanka only a few areas are catered by the national educational institutes [1]. By finding out the educational desires of the youth, it will be possible to design and develop educational and professional programmes and institutes which can be readily accepted by the youth and give better results than that can be achieved by only pursuing traditional programmes. Data Mining which is a powerful tool that can recognize and unearth significant facts, relationships, trends and patterns can be employed to discover this information [2]. In this project, a data mining model has been developed to predict the educational desire of youths at an early stage from other social data.

II. THEORETICAL BACKGROUND

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data [3]. Univariate analysis is the simplest form of quantitative (statistical) analysis.

The analysis is carried out with the description of a single variable and its attributes of the applicable unit of analysis. Univariate analysis contrasts with bivariate analysis – the analysis of two variables simultaneously – or multivariate analysis – the analysis of multiple variables simultaneously. Univariate analysis is also used primarily for descriptive purposes, while bivariate and multivariate analysis are geared more towards explanatory purposes [4]. Univariate analysis is commonly used in the first stages of research, in analyzing the data at hand, before being supplemented by more advance, inferential bivariate or multivariate analysis [5]. Pearson's chi-square test is the best-known of several chi-square tests – statistical procedures whose results are evaluated by reference to the chi-square distribution [6].

With large samples, a chi-square test can be used. However, the significance value it provides is only an approximation, because the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution. The approximation is inadequate when sample sizes are small, or the data are very unequally distributed among the cells of the table, resulting in the cell counts predicted on the null hypothesis (the "expected values")
being low. The usual rule of thumb for deciding whether the chi-squared approximation is good enough is that the chi-squared test is not suitable when the expected values in any of the cells of a contingency table are below 5 or below 10 when there is only one degree of freedom [7]. In contrast the Fisher exact test is, as its name states, exact, and it can therefore be used regardless of the sample characteristics [8]. For hand calculations, the test is only feasible in the case of a 2×2 contingency table. However the principle of the test can be extended to the general case of an m×n table [9].

Logistic regression is most frequently employed to model the relationship between a dichotomous (binary) outcome variable and a set of covariants, but with a few modifications it may also be used when the outcome variable is polytomous [10].

The extension of the model and the methods from a binary outcome variable to a polytomous outcome variable can be easily illustrated when the outcome variable has three categories. Further generalization to an outcome variable with more than three categories is more of a notation problem than a conceptual one [11]. Hence, it will be considered only the situation when the outcome variable has three categories.

Main objective of fitting this statistical model is to find out the sequence of variables being significant to the model, so that the sequence of variables, as a whole or a subsequence starting from the first variable, will be used as necessary in constructing a decision tree. In this study we make use of this statistical model not for interpretation but only for doing a comparison with the outcome of a Data Mining approach in decision making.

III. ANALYSIS

Univariate analysis is carried out with the purpose of analyzing each variable independently from other variables. Therefore each of the categorical variables measured as a factor is cross tabularized with the dependent variable “Type of Further Educational Desires” calculating percentages of respondents belonging to each combination of levels, and the Chi-Square Test is performed in order to measure the strength of association between factors and the response of interest. A tolerance rate of 20 percent has been fixed as the significance level for further analysis. Table 1 shows the results of the analysis.

| Explanatory Variable | Pearson’s Chi-square Value | Degrees of Freedom | Asymptotic Level of Significance | Significant Variables at 20% level |
|----------------------|----------------------------|--------------------|---------------------------------|----------------------------------|
| Age group            | 208.0                      | 6                  | 0.0000                          | √                                |
| Gender               | 68.751                     | 3                  | 0.0000                          | √                                |
| Educational level    | 198.1                      | 6                  | 0.0000                          | √                                |
| Ethnicity            | 53.032                     | 12                 | 0.0000                          | √                                |
| Province             | 198.1                      | 24                 | 0.0000                          | √                                |
| Sector               | 6.882                      | 6                  | 0.2020                          |                                   |
| Social class (self definition) | 77.154                     | 9                  | 0.0000                          | √                                |
| Present financial    | 25.455                     | 6                  | 0.0000                          | √                                |

From the results shown in Table 1, the variables with their p-values less than 0.2 have been detected as significant.

A. Fitting a Statistical Model

The main purpose of this part of analysis is to determine the factors, which affect or are associated with having different types of Further Educational Desires in youth. Though several variables have been identified as significant factors, where each could independently build a significant effect on developing different wishes on education among youth, due to the confounding nature of these factors, it is not easy to conclude on their corporate influence on making Further Education Desires different in people.

Therefore this modeling approach can be very much useful in detecting the genuine effect of these factors when adjusted for some other factors as well.

Since the response variable is Multinomial and the scale of response levels are Nominal, it was decided to work out a “logit” link in regression modeling. Therefore a Generalized Logit Model will be fitted to accomplish the objective.

B. Fitting the Best Fitted Generalized Logit Model

The Forward Selection procedure is used in selecting variables to the model. In assessing the fit of the terms to the model, the difference in deviance of the two models compared, which is distributed as Chi-Squared has been used at the 5% significance level. However the terms will be selected to the model, as they do the best representation of all the data.

The results obtained in following the steps of fitting a Generalized Logit Model using the procedure CATMOD in SAS package, are tabularized in the body of the analysis.

Let the Null Model

\[ \log \left( \frac{\pi_f}{\pi_0} \right) = \alpha_i; \]

where \( \pi_f \) is the probability that a respondent has the \( f^{th} \) type of Further Education Desire, \( f \neq F \), and type \( F \) is the Further Education Desire category “No Desire”.

Fitting Main Effects to the Model

Step 1: Null Model vs One Variable Model (Model 1)

Table 2 shows sample of data used in devising Model 1.
Table 2: Adding the 1st Most Significant Variable to the Model

| Model 1 (terms added) | Raw Deviance | Difference in Deviance | Difference in df | p-value |
|-----------------------|--------------|------------------------|------------------|---------|
| Null Model            | 4540.468     | 0                      | -                | -       |
| Age group             | 4332.934     | 207.534                | 4                | 9.00999E-44 |
| Gender                | 4474.573     | 65.8957                | 2                | 4.90829E-15 |
| Educational level     | 4359.197     | 181.2715               | 4                | 3.97611E-38 |
| Ethnicity             | 4490.291     | 50.1769                | 7                | 1.33346E-08 |
| Province              | 4401.342     | 139.1265               | 16               | 1.06808E-21 |
| Social class          | 4477.97      | 62.4984                | 6                | 1.39659E-11 |
| Financial situation   | 4517.147     | 23.3216                | 4                | 0.000109204 |
| Major problems with education | 4452.271 | 88.1975               | 14               | 8.30242E-13 |
| Access to educational facilities | 4450.442 | 90.0261               | 4                | 1.30006E-18 |
| Type of activity      | 3450.942     | 1089.527               | 10               | 9.5521E-228 |

From Table 2, the lowest p-value is associated with the variable ‘Type of activity’. The selection procedure of the most significant variable requires that ‘Type of activity’ be added to the Null Model as the first step of developing a model where the ‘Type of Further Education Desire’ being the response variable.

The explanatory variables in the model: Type of activity

Model 1: $\log \frac{\pi_{ij}}{\pi_{ij}} = \alpha_i + x_i \beta_f$

where $\pi_{ij}$ is the probability that a respondent in ‘Type of activity’ category i has the fth type of Further Education Desire, $f \neq F$, and type F is the Further Education Desire category “No Desire”.

Thus two logits are modeled for each activity type: the logit comparing Technical/Vocational Education to No Desire and the logit comparing University/Higher Education to No Desire.

Model 1: for Type of activity i

logit(response1/response3)i=1 models the probability of response category 1 relative to the response category 3.

logit(response2/response3)i=2 models the probability of response category 2 relative to the response category 3.

There are separate sets of intercept parameters $\alpha_i$ and regression parameter $\beta_f$ for each logit and the matrix $x_i$ is the explanatory variable for the $i^{th}$ population.

Step 2: Model 1 vs Two Variable Model (Model 2)

Table 3 shows sample of data used in devising Model 2.

Table 3: Adding the 2nd Most Significant Variable to the Model

| Model 2 (terms added) | Raw Deviance | Difference in Deviance | Difference in df | p-value |
|-----------------------|--------------|------------------------|------------------|---------|
| Age group             | 4337.718     | 13.2236                | 4                | 0.01023388 |
| Gender                | 3434.514     | 16.4276                | 2                | 0.000270889 |
| Educational level     | 3288.146     | 162.7954               | 4                | 3.67573E-34 |
| Ethnicity             | 3427.465     | 23.4763                | 7                | 0.001407636 |
| Province              | 3316.505     | 134.4366               | 16               | 8.79825E-21 |
| Social class          | 3420.852     | 30.0896                | 6                | 3.77963E-05 |
| Financial situation   | 3432.545     | 18.3971                | 4                | 0.001031951 |
| Financial situation in past | 3448.478 | 2.4638                | 4                | 0.65112949 |
| Major problems with education | 3427.67 | 23.2713               | 14               | 0.055995893 |
| Access to educational facilities | 3437.028 | 13.9133               | 4                | 0.00757697 |

Since the variable ‘Educational Level’ has the lowest p-value, it was brought into the model that has been adjusted for ‘Type of activity’.

The explanatory variables in the model: Type of activity, Educational Level

Model 2: $\log \frac{\pi_{ij}}{\pi_{ij}} = \alpha_i + x_i \beta_f$

where the matrix $x_i$ is the set of explanatory variables for the $i^{th}$ population.

Step 3: Model 2 vs Three Variable Model (Model 3)

Table 4 shows sample of data used in devising Model 3.

Table 4: Adding the 3rd Most Significant Variable to the Model

| Model 3 (terms added) | Raw Deviance | Difference in Deviance | Difference in df | p-value |
|-----------------------|--------------|------------------------|------------------|---------|
| Age group             | 3276.062     | 12.0842                | 4                | 0.016735994 |
| Gender                | 3268.712     | 19.4339                | 2                | 6.02535E-05 |
| Ethnicity             | 3267.805     | 20.3416                | 7                | 0.00487722 |
| Province              | 3155.297     | 132.8494               | 16               | 1.79279E-20 |
| Social class          | 3266.442     | 21.7041                | 6                | 0.001369785 |
| Financial situation   | 3276.61      | 11.536                 | 4                | 0.02115679 |
| Financial situation in past | 3285.776 | 2.3706                | 4                | 0.667946717 |
| Major problems with education | 3267.495 | 20.6513               | 14               | 0.110907989 |
| Access to educational facilities | 3281.422 | 6.7239                | 4                | 0.151218299 |

Since the variable ‘Province’ has the lowest p-value, it was brought into the model already adjusted for Type of activity and Educational Level.

www.ijacsa.thesai.org
The explanatory variables in the model: Type of activity, Educational Level, Province

Model 3: \[ \log \left( \frac{\pi_{ijk}}{\pi_{ij}} \right) = \alpha_I + \chi_{ijk} \beta_I \]

where the matrix \( X_{ijk} \) is the set of explanatory variables for the \( ijk^{th} \) population.

Step 4: Model 3 vs Four Variable Model (Model 4)
Table 5 shows sample of data used in devising Model 4.

Table 5: Adding the 4th Most Significant Variable to the Model

| Model 4 (terms added) | Raw Deviance (-2Log likelihood) | Difference in Deviance | Difference in df | p-value |
|-----------------------|---------------------------------|------------------------|-----------------|---------|
| Model 3               | 3155.297                        | 0                      | -               | -       |
| Age group             | 3145.734                        | 9.5632                 | 4               | 0.048464702 |
| Gender                | 3133.988                        | 21.3087                | 2               | 2.3598E-05  |
| Ethnicity             | 3152.673                        | 2.6242                 | 7               | 0.917458009 |
| Social class          | 3128.48                         | 26.8167                | 6               | 0.000156716 |
| Present financial situation | 3146.391                     | 8.9055                 | 4               | 0.063505426 |
| Financial situation in past | 3152.922                    | 2.3747                 | 4               | 0.667204138 |
| Major problems with education | 3138.381                  | 16.9154                | 14              | 0.260714267 |
| Access to educational facilities | 3140.936                  | 14.3609                | 4               | 0.006227982 |

In Step 4, Gender was found significant and brought into the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender

Model 4: \[ \log \left( \frac{\pi_{ijk}}{\pi_{ij}} \right) = \alpha_I + \chi_{ijk} \beta_I \]

where the matrix \( X_{ijk} \) is the set of explanatory variables for the \( ijk^{th} \) population.

Step 5: Model 4 vs Five Variable Model (Model 5)
Table 6 shows sample of data used in devising Model 5.

Table 6: Adding the 5th Most Significant Variable to the Model

| Model 5 (terms added) | Raw Deviance (-2Log likelihood) | Difference in Deviance | Difference in df | p-value |
|-----------------------|---------------------------------|------------------------|-----------------|---------|
| Model 4               | 3133.988                        | 0                      | -               | -       |
| Age group             | 3124.457                        | 9.5309                 | 4               | 0.04911679 |
| Ethnicity             | 3131.679                        | 2.3095                 | 7               | 0.940746591 |
| Social class          | 3108.071                        | 25.9174                | 6               | 0.00023067 |
| Present financial situation | 3125.265                    | 8.7235                 | 4               | 0.068394724 |
| Financial situation in past | 3131.218                   | 2.7702                 | 4               | 0.596987596 |
| Major problems with education | 3117.493                  | 16.495                 | 14              | 0.284088473 |

In Step 5, ‘Social class’ was found significant and brought into the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender, Social class

Model 5: \[ \log \left( \frac{\pi_{ijklm}}{\pi_{ijkm}} \right) = \alpha_I + \chi_{ijklm} \beta_I \]

where the matrix \( X_{ijklm} \) is the set of explanatory variables for the \( iklmn^{th} \) population.

Step 6: Model 5 vs Six Variable Model (Model 6)
Table 7 shows sample of data used in devising Model 6.

Table 7: Adding the 6th Most Significant Variable to the Model

| Model 6 (terms added) | Raw Deviance (-2Log likelihood) | Difference in Deviance | Difference in df | p-value |
|-----------------------|---------------------------------|------------------------|-----------------|---------|
| Model 5               | 3133.988                        | 0                      | -               | -       |
| Age group             | 3124.457                        | 9.5309                 | 4               | 0.04911679 |
| Ethnicity             | 3131.679                        | 2.3095                 | 7               | 0.940746591 |
| Social class          | 3108.071                        | 25.9174                | 6               | 0.00023067 |
| Present financial situation | 3125.265                    | 8.7235                 | 4               | 0.068394724 |
| Financial situation in past | 3131.218                   | 2.7702                 | 4               | 0.596987596 |
| Major problems with education | 3117.493                  | 16.495                 | 14              | 0.284088473 |

In Step 6, ‘Age group’ was found significant and brought into the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender, Social class, Age group

Model 6: \[ \log \left( \frac{\pi_{ijklmn}}{\pi_{ijklmn}} \right) = \alpha_I + \chi_{ijklmn} \beta_I \]

where the matrix \( X_{ijklmn} \) is the set of explanatory variables for the \( iklmn^{th} \) population.

It was observed that addition of the remaining variables did not improve the results. Hence the Model 6 has been identified as the best main effect model.

C. Improving the Model

It was further investigated to determine if the addition of two way interaction terms improved the model. The importance of an interaction term was assessed by checking the impact of the difference in deviance of the model.

Step 7: Model 6 Vs Model 7
Table 8 shows sample of data used in devising Model 7.
The interaction between ‘Education Level’ and Province has been found to have an effect on the model and hence added to the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender, Social class, Age group, Edu*Pro, Finp*Mprob

Model 8:

\[
\log \left( \frac{\pi_{ijklmn}}{\pi_{ijklmnF}} \right) = \alpha + x_{ijklmn} \beta + y_{(jk)F} \beta + z_{(op)} \beta + t_{(kl)F} \beta
\]

where the matrix \(x_{ijklmn}\) is the set of explanatory variables for the \(ijklmn^{th}\) population, the matrix \(y_{(jk)}\) is the interaction effect of Edu*Pro and the matrix \(z_{(op)}\) is the interaction effect of Finp*Mprob.

Step 9: Model 8 Vs Model 9

Table 10 shows sample of data used in devising Model 9.

The interaction between ‘Major Problems with Education (Mprob)’ and ‘Province (Pro)’ has been found to have an effect on the model and hence added to the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender, Social class, Age group, Edu*Pro, Finp*Mprob, Mprob*Pro

Model 9:

\[
\log \left( \frac{\pi_{ijklmn}}{\pi_{ijklmnF}} \right) = \alpha + x_{ijklmn} \beta + y_{(jk)F} \beta + z_{(op)} \beta + t_{(kl)F} \beta + u_{(ki)F} \beta
\]

Table 10: Adding the 3rd Most Significant 2-Way Variable to the Model

In this step, the interaction between ‘Major Problems with Education (Mprob)’ and ‘Province (Pro)’ has been found to have an effect on the model and hence added to the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender, Social class, Age group, Edu*Pro, Finp*Mprob, Mprob*Pro

Model 9:

\[
\log \left( \frac{\pi_{ijklmn}}{\pi_{ijklmnF}} \right) = \alpha + x_{ijklmn} \beta + y_{(jk)F} \beta + z_{(op)} \beta + u_{(ki)F} \beta
\]

In this step, the interaction between ‘Major Problems with Education (Mprob)’ and ‘Province (Pro)’ has been found to have an effect on the model and hence added to the model.

The explanatory variables in the model: Type of activity, Educational Level, Province, Gender, Social class, Age group, Edu*Pro, Finp*Mprob, Mprob*Pro

Table 9: Adding the 2nd Most Significant 2-Way Variable to the Model

Table 9 shows sample of data used in devising Model 8.

Table 8: Adding the 1st Most Significant 2-Way Variable to the Model

Table 10 shows sample of data used in devising Model 9.
where the matrix $x_{ijklmn}$ is the set of explanatory variables for the $ijklmn$th population, the matrix $y_{(jk)}$ is the interaction effect of Edu*Pro, the matrix $z_{(op)}$ is the interaction effect of Finp*Mprob and the matrix $t_{(ko)}$ is the interaction effect of Mprob*Pro.

After Step 9 no more significant two-way terms were revealed.

D. Checking the Adequacy of the Best 2-Way Interaction Model

Goodness of Fit: Hypothesis Testing with Deviance Statistics

$H_0$: No lack of fit

$H_1$: There is some lack of fit

Table 11: Goodness of Fit Testing

|                | Chi-Square | df | p-value |
|----------------|------------|----|---------|
| Deviance       | 2145.881   | 2930 | 1.000   |

According to the p-value, it can be concluded that the Null Hypothesis is not rejected. Hence, the test is the proof that there is no lack of fit of the model or the model developed fits the data well.

E. Classification Table

The classification table was used to evaluate the predictive accuracy of the regression model.

Table 12: Classification Table for Model 9

|                | NO DESIRE | TECHNICAL/ VOCATIONAL EDUCATION | UNIVERSITY/ HIGHER EDUCATION | % CORRECT |
|----------------|-----------|--------------------------------|-----------------------------|-----------|
| NO DESIRE      | 641       | 141                            | 51                          | 77.0%     |
| TECHNICAL/ VOCATIONAL EDUCATION | 249   | 304                            | 89                          | 47.4%     |
| UNIVERSITY HIGHER EDUCATION     | 46       | 45                             | 519                         | 85.1%     |
| OVERALL, PERCENTAGE             | 44.9%    | 23.5%                          | 31.6%                       | 70.2%     |

From the results shown in Table 12, it can be seen that Model 9 has the accuracy of more than 70 percent.

F. Implementation of Data Mining Techniques

Finally the model (Model 9) developed using Univariate analysis was used to develop a data mining model. This data mining model can predict the ‘Further Educational Desire’ in a youth from other attributes discussed above.

Construction of Decision Tree

A decision tree was constructed using the attributes identified as significant in the Univariate analysis. The ordering of attributes in the decision tree was also as determined in the statistical analysis. Table 13 shows the attributes and the no. of levels of each attribute. ‘Type of Further Education Desire’ is the class attribute.

Table 13: List of Attributes used in the Decision Tree

| Seq. No | Attribute Name          | No. of Levels |
|---------|-------------------------|---------------|
| 1       | Type of Activity        | 7             |
| 2       | Educational Level       | 3             |
| 3       | Province                | 9             |
| 4       | Gender                  | 2             |
| 5       | Social Class            | 4             |
| 6       | Age group               | 3             |
| 7       | Financial Situation in Past | 3         |
| 8       | Major Problems with Education | 8       |
| 9       | Type of Further Education Desire | 3     |

Figure 1 shows the portion of the decision tree thus constructed.

Figure 2 shows a sample classification rule set developed using the Decision Tree shown in Figure 1.
Rule 1: Type of Activity=Permanently Employed ^ Educational Level=No Schooling/Grade 1-5 ^ Province=Western ^ Gender=Male ^ Social Class=Middle Class ^ Age Group=20-24 yrs
   No Desire
Rule 2: Type of Activity=Permanently Employed ^ Educational Level=No Schooling/Grade 1-5 ^ Province=Western ^ Gender=Male ^ Social Class=Working Class ^ Age Group=15-19 yrs
   No Desire
Rule 3: Type of Activity=Permanently Employed ^ Educational Level=No Schooling/Grade 1-5 ^ Province=Central ^ Gender=Male ^ Social Class=Working Class ^ Age Group=20-24 yrs
   No Desire
Rule 4: Type of Activity=Permanently Employed ^ Educational Level=No Schooling/Grade 1-5 ^ Province=Central ^ Gender=Female ^ Social Class=Working Class ^ Age Group=20-24 yrs
   No Desire
Rule 5: Type of Activity=Permanently Employed ^ Educational Level=No Schooling/Grade 1-5 ^ Province=Uva ^ Gender=Male ^ Social Class=Working Class ^ Age Group=24-29 yrs
   Technical/Vocational Education
Figure 2: Sample Rule Set

G. Implementation
A software tool was developed using Visual Basic to implement the rule set developed above. Figure 3 shows the interface of the software tool developed.

Figure 3: Interface of the Software Tool Developed

H. Evaluation
The system developed was tested using a random set of data containing 485 test records. The selected random set was divided into four test sets each test set containing around 125 records. Table 14 shows the classification table obtained by inputting data into the application.

Table 14: Classification Table Obtained from Test

| OBSERVED          | PREDICTED                  |
|-------------------|----------------------------|
|                  | TECHNICAL/VOCATIONAL EDUCATION | UNIVERSITY/HIGHER EDUCATION | NO DESIRE |
| TECHNICAL/VOCATIONAL EDUCATION | 21 | 13 | 19 | 22 | 3 | 5 | 5 | 7 | 13 | 17 | 16 | 7 |
| UNIVERSITY/HIGHER EDUCATION | 0  | 4  | 3  | 7  | 30 | 23 | 35 | 26 | 3  | 5  | 2  | 2 |
| NO DESIRE         | 10 | 5  | 11 | 8  | 4  | 7  | 2  | 0  | 32 | 51 | 27 | 40 |

Table 15 shows the Confusion Matrices for various adjusted outcomes.

Table 15 (a): Confusion Matrix Adjusted for No Desire

| Observed | No Desire | Have a Desire |
|----------|-----------|---------------|
| Data Set 1 | No Desire | 36 | 14 |
|          | Have a Desire | 16 | 54 |
| Data Set 2 | No Desire | 51 | 12 |
|          | Have a Desire | 22 | 45 |
| Data Set 3 | No Desire | 27 | 13 |
|          | Have a Desire | 18 | 62 |
| Data Set 4 | No Desire | 40 | 8 |
|          | Have a Desire | 9  | 62 |

Table 15 (b): Confusion Matrix Adjusted for Uni/Higher Education

| Observed         | Uni/Higher | Not So |
|------------------|------------|--------|
| Data Set 1       | No Desire | 30  | 3   |
|                  | Have a Desire | 7  | 76  |
| Data Set 2       | No Desire | 23  | 9   |
|                  | Have a Desire | 12 | 86  |
| Data Set 3       | No Desire | 35  | 5   |
|                  | Have a Desire | 7  | 73  |
| Data Set 4       | No Desire | 26  | 9   |
|                  | Have a Desire | 7  | 77  |

Table 15 (c): Confusion Matrix Adjusted for Tech/Voc. Education

| Observed        | Tech/Voc | Not So |
|-----------------|----------|--------|
| Data Set 1      | No Desire | 21  | 16  |
|                  | Have a Desire | 10 | 69  |
| Data Set 2      | No Desire | 13  | 22  |
|                  | Have a Desire | 9  | 86  |
| Data Set 3      | No Desire | 19  | 21  |
|                  | Have a Desire | 14 | 66  |
| Data Set 4      | No Desire | 22  | 14  |
|                  | Have a Desire | 15 | 68  |

Table 16 shows the resulting measures obtained from the tests.

Table 16: Measures Obtained from Tests

|                   | Tech/Voc | Uni/High | No Desire | Avg. |
|-------------------|----------|----------|-----------|------|
| Data Set 1        | TPR      | FPR      | Accuracy  |
|                   | 0.567568 | 0.126582 | 0.775862  |
|                   | Precision | Recall | FMeasure |
|                   | 0.75051  | 0.14426  | 0.77182   |
| Data Set 2        | TPR      | FPR      | Accuracy  |
|                   | 0.371429 | 0.094737 | 0.761538  |
|                   | Precision | Recall | FMeasure |
|                   | 0.75946  | 0.1818   | 0.738462  |
| Data Set 3        | TPR      | FPR      | Accuracy  |
|                   | 0.761538 | 0.094737 | 0.738462  |
|                   | Precision | Recall | FMeasure |
|                   | 0.77182  | 0.1818   | 0.738462  |
| Data Set 4        | TPR      | FPR      | Accuracy  |
|                   | 0.611111 | 0.180723 | 0.703514  |
|                   | Precision | Recall | FMeasure |
|                   | 0.703514 | 0.180723 | 0.703514  |
| Overall           | TPR      | FPR      | Accuracy  |
|                   | 0.75051  | 0.14426  | 0.77182   |
|                   | Precision | Recall | FMeasure |
|                   | 0.77182  | 0.1818   | 0.77182   |

From Table 16, it can be seen that the overall accuracy of the system is above 80 percent. Figure 7 shows the Receiver-Operator Characteristic (ROC) curve drawn from the data in Figure 4.
From the ROC curve, it can be seen that the results concentrate towards the upper left hand corner. This is the proof that the accuracy of the Data Mining model is acceptable as pure random guess would lie along the diagonal line.

IV. CONCLUSIONS

This paper presented the results of the research carried out to find the factors on which the Educational Desires of Sri Lankan Youth depends. The research found that the Educational Desires of the Youth could be predicated through the combination of several social factors. The findings of the research were finally used to design a data mining model for the predication of the Educational Desires of the Youth. This model can be used by decision makers in dealing with issues concerning youth especially their further educational requirements.

REFERENCES

[1] Hettige ST, Mayer M, “Sri Lankan Youth – Challenges and Responses”, Friedrich Ebert Stiftung, Colombo, Sri Lanka, (2002).
[2] Fernando HC, Tissera WMR, Athauda RI, “Gaining Insights to the Tea Industry of Sri Lanka using Data Mining”, Proceedings of the International Multi Conference of Engineers and Computer Scientists (IMECS 2008) Vol I, Hong Kong, (2008).
[3] Research Methods Knowledge Base, “Descriptive Statistics”, http://www.socialresearchmethods.net/kb/statdesc.php [Accessed on 08/04/2010]
[4] Babbie ER, “The Practice of Social Research”, 12th edition, Wadsworth Publishing, (2009).
[5] Cooper RA, Weekes AJ, “Data, Models, and Statistical Analysis”, Rowman & Littlefield Publishers, Inc., USA, (1983).
[6] Wikipedia, “Pearson's Chi-Square Test”, http://en.wikipedia.org/wiki/Pearson%27s_chi-square_test [Accessed on 08/04/2010]
[7] Kinley L, “Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics”, Journal of the American Statistical Association 73 (362): pp. 253–263, (1978).
[8] Mehta CR, Patel NR, Tsatis AA, "Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data". Biometrics 40 (3): pp. 819–825, (1984).
[9] Weisstein EW, "Fisher's Exact Test." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/FishersExactTest.html [Accessed on 08/04/2010]
[10] Sarkar SK, Midi H, “Optimization Techniques for Variable Selection in Binary Logistic Regression Model Applied to Desire for Children Data” Journal of Mathematics and Statistics 5 (4): pp. 387-394, (2009).
[11] Pahwa P, Karunanayake CP, “Modeling of Longitudinal Polytomous Outcome from Complex Survey Data - Application to Investigate an Association between Mental Distress and Non-Malignant Respiratory Diseases”, BMC Medical Research Methodology, Vol 9, (2009).

AUTHORS PROFILE

Mohamed Fazil Mohamed Firdhous is a senior lecturer attached to the Faculty of Information Technology of the University of Moratuwa, Sri Lanka. He received his BSc Eng., MSc and MBA degrees from the University of Moratuwa, Sri Lanka, Nanyang Technological University, Singapore and University of Colombo Sri Lanka respectively. In addition to his academic qualifications, he is a Chartered Engineer and a Corporate Member of the Institution of Engineers, Sri Lanka, the Institution of Engineering and Technology, United Kingdom and the International Association of Engineers. Mohamed Firdhous has several years of industry, academic and research experience in Sri Lanka, Singapore and the United States of America.

Ravindi Jayasundara is a lecturer attached to the Department of Mathematics, Faculty of Engineering of the University of Moratuwa, Sri Lanka. She completed the BSc (Hons) degree in Statistics with Second Class (Upper Division) at the University of Colombo, Sri Lanka and MSc in Operations Research at the University of Moratuwa, Sri Lanka. Ravindi’s research interests are in statistics, operations research and data mining.