International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Gender Prediction on Scholarship Grant using Supervised Learning Techniques

Emmer P. Ruaya*

Assistant Professor II, North Eastern Mindanao State University- Tagbina Campus

ABSTRACT

Through the advent of the technology, a number of electronic data are being stored and number of transaction is rapidly increased. Using various machine learning techniques it is now possible to collect patterns of knowledge base transforming into valuable information that can be used to predict future decisions. This paper utilized the supervised learning technique to extract the relevant information from existing data sets collected for scholarship grant. The study employs an open source machine learning software called Waikato Environment for Knowledge Analysis (WEKA). Discussed the method in constructing the model and the recognition of its accuracy rate using the classification algorithm Logistic Regression on gender prediction on scholarship grants.

Keywords: Dataset, General Weighted Average, Classification, Logistic Regression, tuition, Gender, Supervised Learning

Introduction

According to the 2016 Global Education Monitoring Report 61 million children do not have access to basic education and 758 million adults in the world are illiterate because they never got any education. Irrespective to different race, ethnicity, gender, religion, age, political preference, nationality and disability has the right to education. Education is the powerful tool on combating poverty, improving socio-economic status, health, equality and peace. Education is the basic right of an individual. Education and poverty inevitably linked because individuals living on poverty may stop on schooling which leaves them to become illiterate and affects to their future children in similar situation. To combat this problem government offers different scholarships grants to offer to the different individuals who are struggling to study. The Commission on Higher Education offers different scholar grants to the different higher education institution to access scholarship grants. However, to due to limited slots of scholarship grants available and number of student are in need. To manage this selection, the Commission sets a meticulous selection so that a deserving student can avail the grant.. Moreover, this study applies supervised learning technique using logistic regression on determining the accuracy on gender scholarship grant; it will enable the commission to make better decision throughout the scholarship grant approval process. In order to achieve this, it is a must to focus on data quality as the quality of decisions to be made depends on it. With accurate data preprocessing, it will improve the quality of the raw experimental data and only allow the selection of variables containing specific and essential information. Since real data frequently contains noise, inconsistencies, redundancies, or even irrelevant information, the elimination of outliers, standardization, and cleaning of data is necessary to achieve the desired quality output.

Review of Literature

This consists of literature survey to prediction of scholarship by using Machine Learning and Data Mining technique.

Suma, V., and ShavigeMalleshwara Hills et al.[1] uses decision trees to assess the student data. Student educational data and information where used as the attributes as dataset and calculated. Classification algorithm where used to identify student performance.

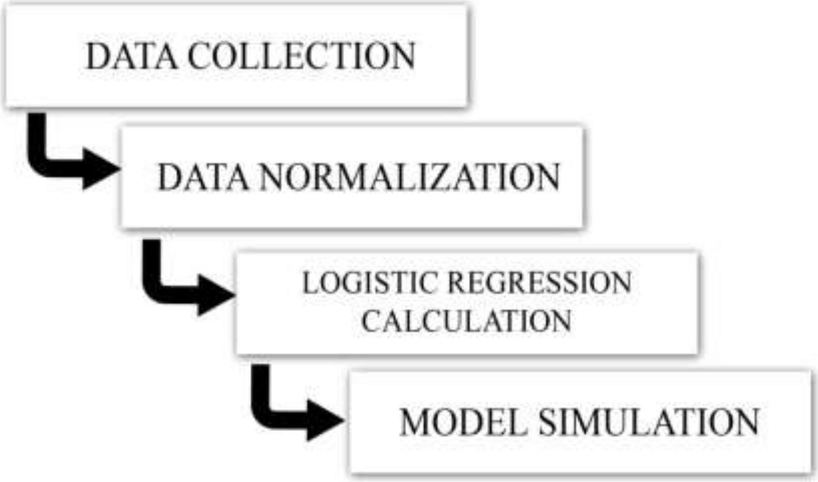
Okfalisa, Ratik a fitriani, YelliVitriana, et al,[5] uses the K-nearest neighbours KNN and linear regression algorithm on predicting. The study compares two models on solving scholarship prediction. It uses Grade point average, statement of letter of active student, Family card, Identity card and result card.

D Kurniadi, E Abdurachman, H L H S Warnars, et al, employs the K-nearest neighbours KNN in analyzing, predicting, and classifying students who have potentials to get scholarships in universities. The attributes used in the prediction process are a semester, parents' income, number of family dependents, and Cumulative Grade Point Average.

Darshana Chaudhari, Rajashri Khadke, et al, comparing the on predicting performance accuracy of two models the Naïve Bayes (NB) and multiclass prediction model using student data on final student grades in the first semester courses.

Methodology

Conceptual research framework. The conceptual framework of research can be seen in Figure 1.



Data Set Collection

The data were used in the study are coming from North Eastern Mindanao State University Tagbina Campus submitted list to Commission on Higher Education and approved list of Tertiary Education Subsidy list of grantees.

Data Normalization

Data normalization is was applied to this study to on cleaning the data organizing the data the appears similarly and eliminating unstructured data and data duplication in order to ensure logical data storage. The data normalization was done using the open source software WEKA by applying the supervised filters to balance the instances of the class.

Logistic Regression

Supervised learning techniques should be considered and used to forecast results on the data provided. The study generates a training set that includes a set of attributes and corresponding output, usually it refers to the target prediction attribute the class. The study aims to discover relationships between the attributes that will require predicted results. Moreover, the Logistic Regression is the best choice for the model, where it has the capability of solving classification problems, and the most common use case is binary. Logistic regression is an open-source Java implementation of the classifiers within the WEKA data mining tool.

Scholarship Forecasting

WEKA an open source software was utilized on the study on analysing the data collected from Eastern Mindanao State University Tagbina Campus. Moreover, Logistic Regression is used on predicting the accuracy of the predicted results.

Results and Discussion

The data collected includes several unnecessary items that might contribute to incomplete analysis; therefore data was preprocessed to satisfy the analytical quality requirements. In this section, the following illustrates the data preprocessing operation that performed using WEKA, see figure 2.

Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

The data were loaded, WEKA recognizes the attributes and, during the data analysis, some basic statistics were calculated for each attribute. Figures below are the results.

Figure 3 Attribute: GWA (General Weighted Average)

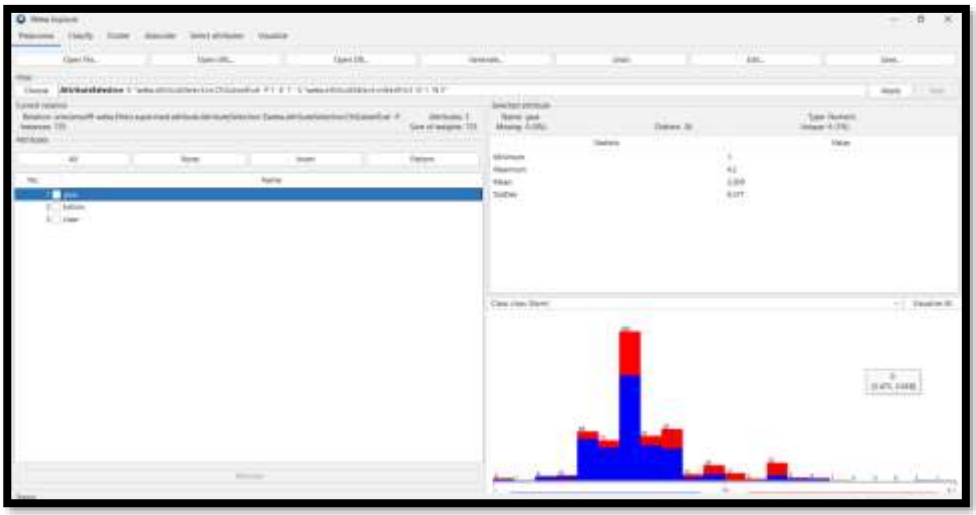


Figure 4 Attribute: Tuition Fee

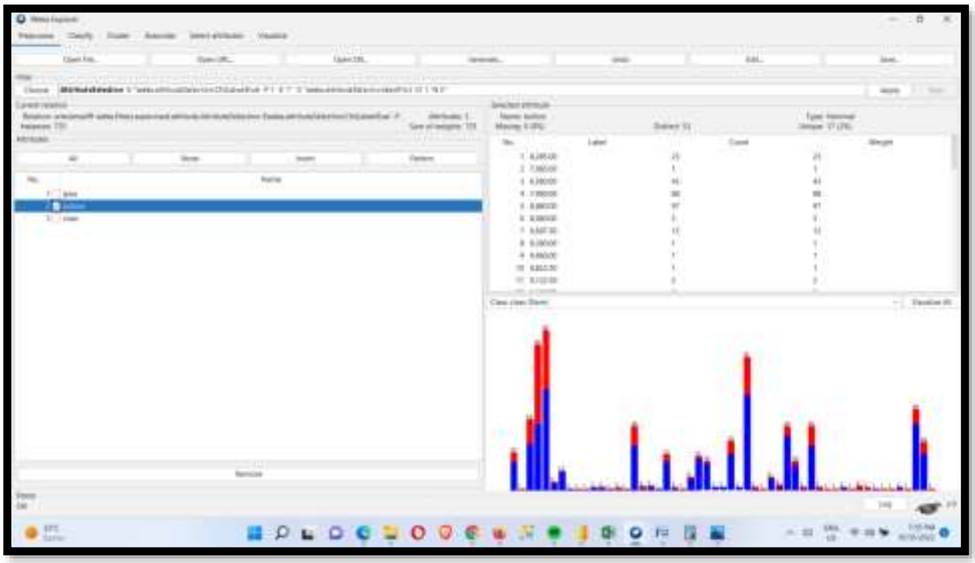


Figure 4 Attribute: Class (Gender)

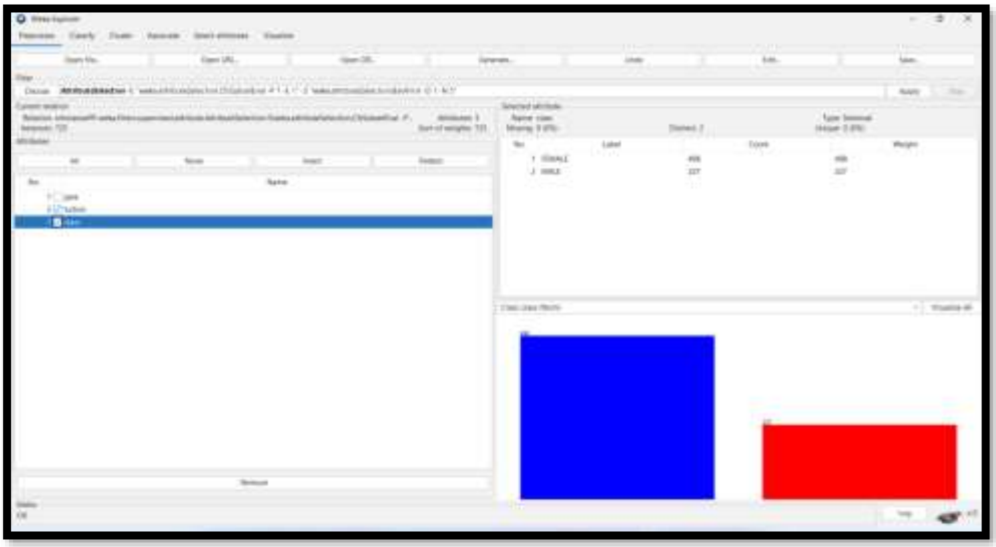
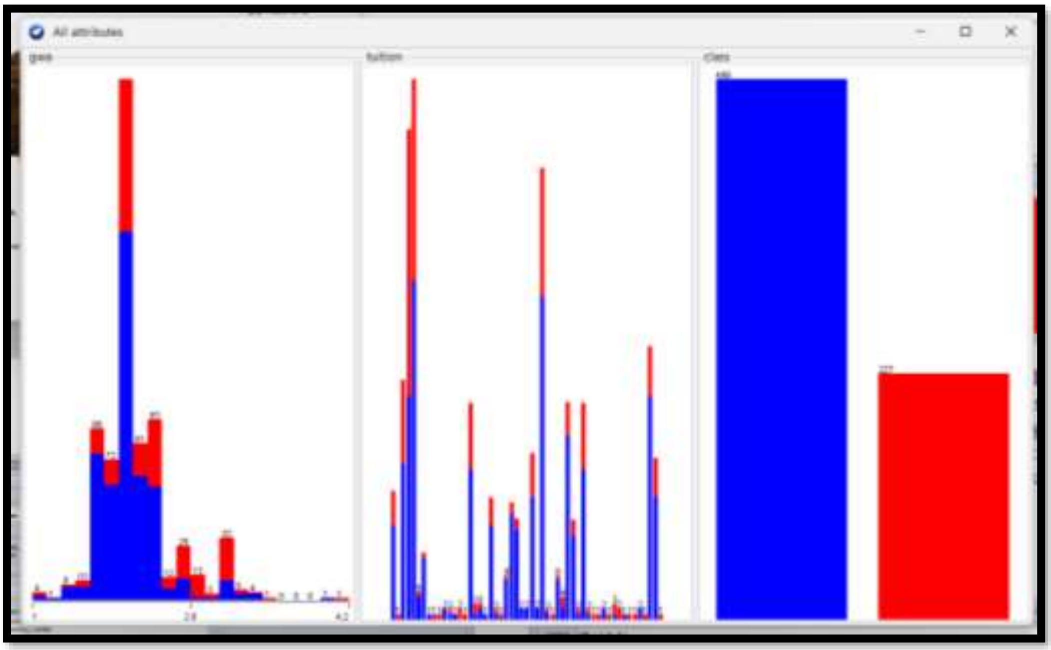
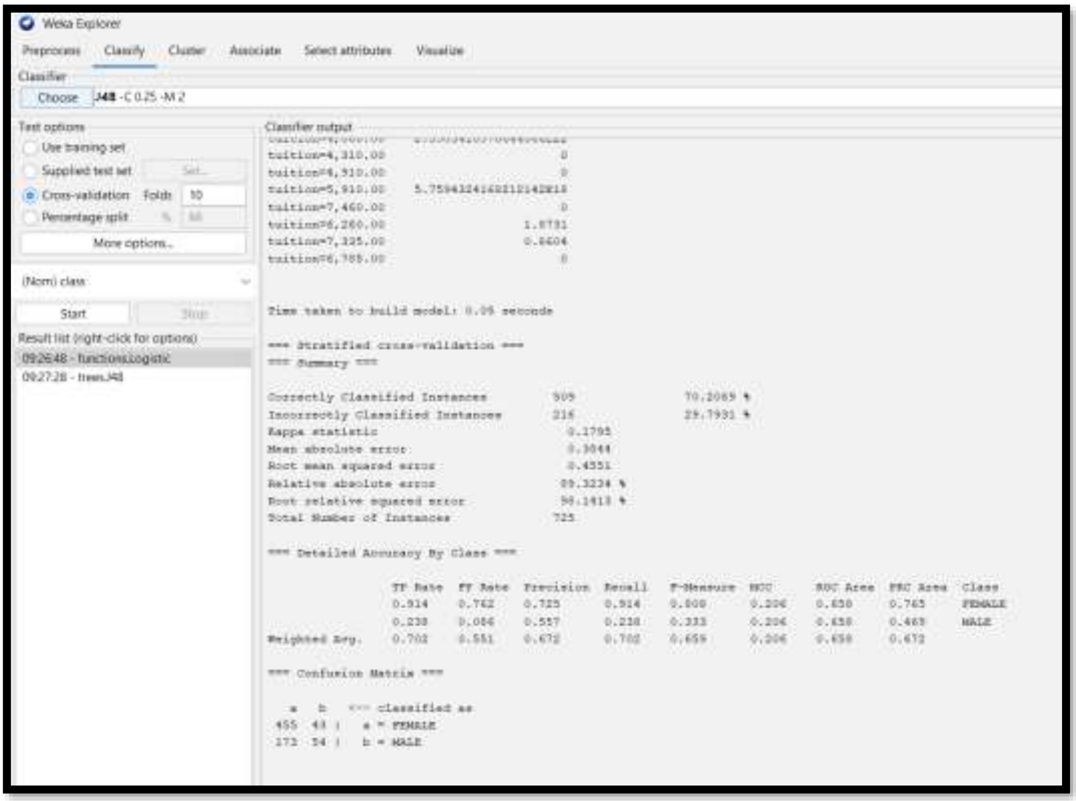


Figure 4 Attribute: Visual Presentation of all attributes



The researcher preferred to use the Logistic regression to develop the model to achieve higher accuracy. The Logistic regression was applied on the dataset witth 725 number of instances and 3 attributes on arff format. This implementation obtained 70.20% of accuracy and build the model 0.05 seconds. The result of the classification is shown on the figure below.



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

09:26:48 - functions.Logistic

09:27:28 - trees.J48

Classifier output

```
tuition=4,310.00                        0
tuition=4,910.00                        0
tuition=5,910.00   5.7594324162212142E18
tuition=7,460.00                        0
tuition=6,260.00                   1.8731
tuition=7,325.00                   0.8604
tuition=6,785.00                        0


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         509               70.2069 %
Incorrectly Classified Instances       216               29.7931 %
Kappa statistic                          0.1795
Mean absolute error                      0.3044
Root mean squared error                  0.4551
Relative absolute error                 89.3224 %
Root relative squared error             98.1813 %
Total Number of Instances              725

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.914    0.762    0.725      0.914    0.808      0.206    0.658     0.765     FEMALE
                 0.238    0.086    0.557      0.238    0.333      0.206    0.658     0.465     MALE
Weighted Avg.    0.702    0.551    0.672      0.702    0.659      0.206    0.658     0.672

=== Confusion Matrix ===

   a   b   <-- classified as
 455  43 |   a = FEMALE
 173  54 |   b = MALE
```

Conclusions

Supervised Learning has been apply on this paper on developing predictive model on gender prediction on granting scholarship that can be used be used for potential for scholarship application comparisons indicating applicants attributes. This also indicates that normalizing or data cleaning plays a vital role in achieving high accuracy rate and the results obtained using the Logistic Regression were highly commendable for indicating a correctly classified instances rate of 70.20%. On pre-processing, patterns outlined in the figures on this paper will be also used for future researches for enhancement.

References

Van der Geer, J., Hanraads, J. A. J., & Lupton, R. A. (2000). The art of writing a scientific article.*Journal of Science Communication, 163*, 51–59.

Strunk, W., Jr., & White, E. B. (1979).*The elements of style* (3rd ed.). New York: MacMillan.

Mettam, G. R., & Adams, L. B. (1999).How to prepare an electronic version of your article. In B. S. Jones & R. Z. Smith (Eds.), *Introduction to the electronic age* (pp. 281–304). New York: E-Publishing Inc.

Fachinger, J., den Exter, M., Grambow, B., Holgerson, S., Landesmann, C., Titov, M., et al. (2004).Behavior of spent HTR fuel elements in aquatic phases of repository host rock formations, 2nd International Topical Meeting on High Temperature Reactor Technology. Beijing, China, paper #B08.

Fachinger, J. (2006). Behavior of HTR fuel elements in aquatic phases of repository host rock formations.*Nuclear Engineering & Design,236*, 54.

Abdallah, E. E., Alzghoul, J. R., & Alzghool, M. (2020). Age and gender prediction in open domain text. *Procedia Computer Science*, *170*, 563-570.

Khan, M. N. A., Qureshi, S. A., & Riaz, N. (2013). Gender classification with decision trees. *Int. J. Signal Process. Image Process. Patt. Recog*, *6*, 165-176.

Kurniadi, D., Abdurachman, E., Warnars, H. L. H. S., & Suparta, W. (2018, November). The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm. In *IOP conference series: materials science and engineering* (Vol. 434, No. 1, p. 012039). IOP Publishing.

Markov, Z., & Russell, I. (2006). An introduction to the WEKA data mining system. *ACM SIGCSE Bulletin*, *38*(3), 367-368.