International Journal of Computer Applications Technology and Research
Volume 11–Issue 04, 91-97, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1104.1002

Classification Algorithm for Career Recommendation System

Robert Masika[1]
Department of Information Technology, School of Computing and Informatics,
Kibabii University,
Bungoma, Kenya.

Dr. Richard Rono[2]
Department of Computer science, School of Computing and Informatics,
Kibabii University,
Bungoma, Kenya.

Dr. Robert Kati[3]
Department of Science and Mathematics Education,
Faculty of Education and Social Sciences,
Kibabii University,
Bungoma, Kenya.

Abstract: The tremendous developments in technology that have been realized in this digital era have greatly improved the way in which data is collected and used in schools. Over the years the number of secondary schools using technology in processing student data has been increasing steadily. As a result, a large amount of data in electronic form has been gathered. Classification algorithms can be used to study the patterns presented in these data and use it to predict a suitable career for a student. In this study classification algorithms were used to predict a suitable career for form four students. The study evaluated the best classification algorithm for implementing the career recommendation system in Kenya. The Cross Industry Standard Process for Data Mining framework was applied to a dataset drawn from form four students in Bungoma County in Kenya. Stratified random sampling was used to select 50 secondary schools and a 10% of candidates were selected from every sampled schools. The collected data were cleansed, preprocessed and analyzed using a data mining tool of RapidMiner. Various classification algorithms were evaluated in predicting a suitable career for a student. The study findings revealed that classification algorithms can be used to predict a suitable career for a student. All the classifiers that were used gave a predictive accuracy of above 88% though deep learning was the most accurate with 97.5%. However, since the classifiers out performed each other in various metrics, therefore using multiple classification algorithms in building the recommendation model can yield better results. The study therefore concludes that classification models comprising of multiple classifiers can be used to predict suitable careers for secondary students.

Keywords: Rapidminer, Classification algorithm, Career choice, Recommendation system, Data mining

1. INTRODUCTION

In life people are constantly faced with situations where they must make choices and sometimes, they do so by basing on insufficient or lack of information of the available alternatives. The quality of the decision made is highly dependent on the available information. A decision on career choice for a secondary school student is so critical since it's the main determinant of their profession and greatly impacts them throughout their life (Mberia and Midigo, 2018). Furthermore, to make an effective decision the student must understand his/her interests and ability as well as have sufficient information about the available career alternatives. However, this isn't always possible since the main source of this information for the students in secondary schools in Kenya is their teachers and parents who in most cases are either unavailable or ill-equipped. As a result, most of them end up making choices which they later regret (Mudulia, 2017). However, the tremendous developments in technology that have been realized in this digital era have greatly improved the way in which such an exercise can be carried out. Currently artificial intelligence techniques are being used in various fields to aid the process of making decisions. One such a technique that is increasingly being used in the education field is Educational Data Mining (EDM).

EDM is an area of research which uses data mining tools and techniques in the analysis of raw data found in the institutions of learning. These tools analyze the various perspectives of the raw data to generate useful information (Han and Kamber, 2006; Raheela et al., 2017). The information may contain hidden patterns and associations which can be helpful in understanding students' environment and making decisions. The knowledge generated from this information can be useful in the validation and evaluation processes of the education systems which can result in improved quality of teaching and learning (Algarni, 2016).

Over the years the number of secondary schools using technology in processing student data has been increasing steadily. As a result, a large amount of data in electronic form has been gathered. Unless processed, this data is poor in information which may result in unreliable decision making. However, this data can be transformed into knowledge that can improve decision making through Knowledge Discovery in Databases (Muhammad and Safawi, 2017; Suhirman et al, 2014).

Many studies have been done on the application of data mining in education. Though, most of these studies focused on predicting learner performance in institutions of higher education. The studies which have dealt with placement of students in various courses and careers include: Yadav and Pal (2017) used decision trees to select students for enrolment in particular courses. Nikita et al (2017) proposed use of RapidMiner mining tool to analyze educational data to suggest career options for high school student in India. Ansari (2017) developed a framework for career selection in Saudi Arabia using nearest neighbor technique. Wabwoba and Mwakondo (2011) proposed the use of trained Artificial Neural networks to select students by Joint admission board for university courses in Kenya. Stephen et al. (2016) used prediction and classification data mining techniques to predict students' enrolment in Science, Technology, Engineering and Mathematics (STEM) courses in higher education in Kenya.

Currently, there is a growing interest in the use of technology to improve decision making in educational institutions. In particular, there is an appreciation that most of the problems experienced in career selection require a new way of addressing them given the lack of an informed and intelligent way to exploit the available data. Therefore, there is an urgent need to identify new methods of analyzing a student's available data with a view of developing an academic profile that can guide in career selection. Thus, it's important to merge expert knowledge and computational techniques to aid in carrying out this exercise. This research aimed at evaluating classification algorithm for career recommendation in education.

2. RELATED WORK

The desire to have technology supported student advising systems accounts for the many studies that have been done in this area. Some of these studies have focused on the course advisory for both undergraduates and post graduate students while others have focused on career guidance for those joining the colleges and universities.

Ansari (2017) presents a framework design combining expert system and data mining techniques to predict a precise career for the scholars. The data used was collected from various public and private high institutions in Saudi Arabia. It comprised 11 career influencing factors contained in scholar's records. These are: Computer skills, office experience, location, financial status, interest, age, parent's culture, job security, medical insurance and children schooling. They used Nearest Neighbor (NN) technique to implement the tool. The main weakness of this framework is that it failed to take in to account some key factors influencing career choice such as student personality, availability of job opportunities and academic achievement.

Nikita et al (2017) designed a web application in Microsoft visual studio that would help high school students to select a course for their career. The system displays three questionnaires and then analyses the response to the questions using decision trees (C5 and C4.5 algorithm) to establish the students' personality, interest and capacity. This result forms the basis for recommendation of the career option by the system. This study took into consideration some key attributes that affect career choice. However, since it was done outside the country it doesn't take in to consideration some unique aspects in Kenyan system.

Elakia et al. (2014) used the RapidMiner data mining tool to show that data mining methods such as classification can be used to analyze student data in order to identify the suitable career options. They collected datasets from response sheets posted on the website based on Holland's self-directed search. The weakness of this system is that it only based the career recommendation on one attribute i.e. student personality.

The study by Yadav and Pal (as quoted by Stephen et al., 2016) developed a tool using ID3 decision tree algorithm to select student to be enrolled in a particular course by evaluating previous student performance. The limitation of this study is that it only considered the previous student performance and left out the other key attributes.

Stephen et al. (2016) used EDM Classification algorithms to predict student enrolment in science, technology, engineering and mathematics (STEM) courses. The study focused on how individual socio-demographic and contextual factors can determine student enrolment in STEM courses in high institutions in Kenya. The limitation of this study is that it only addressed the issue of learner enrolment in STEM causes.

Wabwoba and Mwakondo (2011) conducted a study on the use of artificial neural networks at Joint Admissions Board in selecting student's university courses for the students to study. The performance data for training was got from JAB and KNEC. They found out that if well trained the ANN application increases the chances of an applicant getting admitted in the career courses in which they are qualified. The limitation of this study is that it only focused on the use of a final exam (KCSE) which may not be representative enough.

In this study a framework that uses classification algorithms to analyze student data and recommend a suitable career for a secondary school student was developed. The study analyzed data collected from three key attributes and used it to build a model that can predict a suitable career for a student.

3. METHODS

The study followed the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology as suggested by Nisbet, Elder and Miner (2009). This methodology has six phases as shown in figure 1 below. It is through these phases that the data mining tool to be used in a real environment is build and implemented.

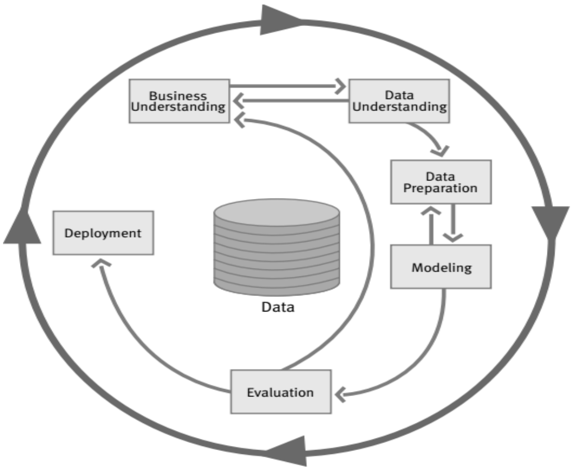

Figure 1: Cross Industry Standard Process for Data Mining

The sub-sections that follow explains in details how the process was used in this study.

3.1 Business understanding phase

This research was developed in the context of secondary schools in Kenya. After admission in form one, a student undertakes a four-year study program and at the tail end selects a career to pursue later. During the admission process and throughout their stay in the school, a lot of student data is usually collected and stored either in the school database, in students' files kept by various departments and/or in the online platforms e.g. NEMIS (national education management information system) etc. However, the collected data may not be in the form which can easily be used therefore this may result in wastage of a precious asset of these institutions.

3.2 Data understanding phase

This phase begun with collecting the raw data from respondents. The collected data captured students' career preferences, their personality and academic achievement. The data was then checked for quality and then scrutinized to get some insights about it to enable formulating the hypothesis that can be used to extract the hidden information (Siraj & Abdoulha, 2009; Stephen et al., 2016).

3.3 Data preparation phase

3.3.1 Data preprocessing

This stage involved preparation of the data for analysis by filling the missing values in the data set, removing irrelevant attributes and selecting relevant ones using feature selection from the full set of attributes. The key attributes included: Sex, English grade, Kiswahili grade, Math grade, Biology grade, Physics grade, Chemistry grade, Religious Education grade, History grade, Geography grade, Technical & Applied grade, student's preference, student's personality and Recommended Career. The missing values in the Recommended Career attribute were manually filled by identifying the best two possible careers for the student based on the collected data. The student's personality was analysed by use of knowledge and rules that are based on Myers-Briggs Typology Indicator (MBTI) and Holland's Self Direct Search (SDS). Then the possible careers were determined from MBTI's Personality Analysis and job suitability Chart and Holland's Vocational Preference Inventory (VPI). These were analysed together with the student's career preferences and their academic achievement to determine the two most preferred careers.

3.3.2 Data transformation

The data was analyzed by a data mining software – RapidMiner. This software was used for implementing this research, since it has a variety of machine learning algorithms that are suitable for carrying out data mining tasks. After selecting the relevant attributes, the data was transformed into an Excel format for ease of analysis since the format (Excel) is acceptable by Rapidminer.

3.4 Modeling phase

The approach was to establish how data mining techniques especially classification techniques can be used to determine whether the selected variables could be used to predict a suitable career for a student. The study used Rapidminer to determine the performance of various classification algorithms based on a number of metrics. Their accuracy levels were evaluated to determine the most effective algorithms that could be used to implement the Recommendation system. The classification algorithms for prediction that were used include; Naïve Bayes, Deep Learning, Decision Tree, Random Forest, k-Nearest Neighbour, and Support Vector Machine. These classifiers are available in Rapidminer tool kit. The following metrics were used to determine the performance of the classification models: time taken to build the model, standard deviation, and prediction accuracy. The total correct predictions given by the algorithm was used to determine its accuracy. The accuracy of the predictive models was calculated based on the percentage of total predictions that were correct. Cross validation evaluation model was used in the final analysis. In this method all the data was divided into 10 disjointed sets of equal sizes. Evaluation was done iteratively such that each of the 10 disjoints took part in the testing and evaluation. Figure 2 shows the classification process.

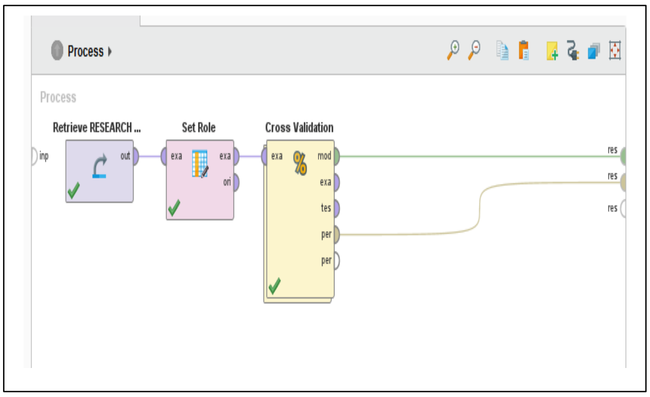

Figure 2: Classification Process

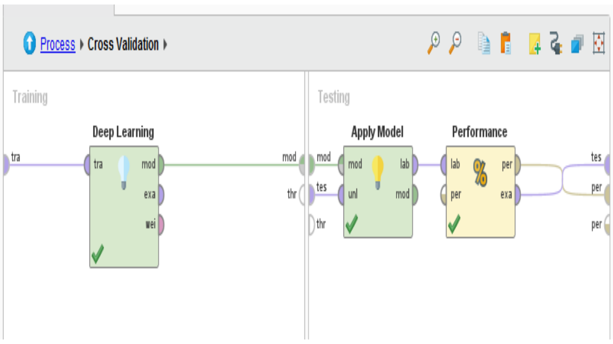

Figure 3.3: cross validation sub process

Figure 3 shows the cross validation sub process

3.5 Evaluation and deployment

This was the final phase of the research design. It involved testing the practical relevance and applicability of the recommender model. An expert survey was performed using focus groups to evaluate the perceived viability of the framework by comparing it with the existing solutions. Descriptive statistics was then be used to analyze these data.

4. RESULTS

The data collected was cleaned to identify the missing and incomplete questionnaires. Of the 455 questionnaires collected only 400 (87.9%) were completely filled. Mugenda (2012) notes that a response rate of 50% or more is adequate. Guided by these thoughts from renowned research academic giants, the response rate for this study was considered to be sufficient in forming conclusions and generalization of the study population.

4.1 Selection of attributes

The following attributes were selected for use in the classification: Sex, English grade, Kiswahili grade, Math grade, Biology grade, Physics grade, Chemistry grade, Religious Education grade, History grade, Geography grade, Technical & Applied grade, Student preference 1, Student preference 2, Student preference 3, SDS personality, MBTI personality, and Recommended Career. The data type of the attribute Sex was nominal while the rest had a data type of poly-nominal. A total of 455 instances were collected from the field.

4.2 Evaluating classification algorithm

The data collected from the students was preprocessed and then used as the input in the Rapidminer data mining tool for processing. Cross validation process was used to divide the data into two sets: training and testing datasets. Cross validation partitioned the data into ten disjoints; nine disjoints served as the training set and one as testing and validation set. The training data set was used to build the models. Six classification algorithms including: Deep Learning, Random Forest, k-Nearest Neighbour, Support Vector Machine, Decision Trees, and Naïve Bayes were used. The models that were obtained from the training data were rerun using the test and validation data sets to evaluate the performance of the resultant models. The process was done iteratively until each of the ten disjoints took part in training as well as testing of the model. The performance of the model was taken as the average of the processes. Six different classifiers were tested using two data sets on 4 parameters. The results are tabulated as shown in Table 1.

Table 1: Performance Measures for the Classifiers for the Career One Dataset

DATASET USED	**CAREER 1**			
CLASSIFIERS USED	MEASURES			
	% ACCURACY	**% ERROR**	**% STD DEV**	**RESPONSE TIME (s)**
Deep Learning	97.6	2.4	±5.0	62
Random Forest	95.6	4.4	±3.2	714
k-Nearest Neighbour	95.6	4.4	±3.2	362
Support Vector Machine	96.6	3.4	±3.2	299
Decision Tree	88.4	11.6	±3.8	59
Naïve Bayes	96.6	3.4	±12.4	54

Table 2: Performance Measures for the Classifiers for the Career Two Dataset

DATASET USED	**CAREER 2**			
CLASSIFIERS USED	MEASURES			
	% ACCURACY	**% ERROR**	**STD DEV**	**TIME (s)**
Deep Learning	87.3	12.7	±2.2	36
Random Forest	85.2	14.8	±2.2	219
k-Nearest Neighbour	86.3	13.7	±2.2	327
Support Vector Machine	87.3	12.7	±2.2	229

Decision Tree	67.6	32.4	±3.2	28
Naïve Bayes	86.5	13.5	±2.2	33

It was noted that most classifiers almost shared the same accuracy, therefore accuracy alone is not the sole determinant in selecting the best classifier. The time taken to build the model, and the standard deviation were also considered in determining the effectiveness of these classifiers.

From table 1, it can be noted Deep Learning classifier achieved highest classification accuracy of 97.6%, with a standard deviation of ±5.0 and response time of 62s. This was closely followed by Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers with a classification accuracy of 96.6%. The standard deviations of SVM and NB were ±3.2% and ±12.4% respectively. SVM had a response time of 299s and NB 54s. Both kNN and Random Forest had an accuracy of 95.6% and a standard deviation of ±3.2%. However, kNN was found to have a better response time of (362s) while random Forest (714s). The classifier that had the lowest accuracy was Decision Tree at 88.4% with a standard deviation of ±3.8 and a response time of 59s.

From Table 2, it's noticed, all classifiers had the same standard deviations of ±2.2%, apart from the Decision Tree classifier (±3.2%). However, they varied in their classification accuracy and response time. Deep Learning (DL) and Support Vector Machine achieved the same classification accuracy of 87.3% but DL had a better response time of 36s as opposed to SVM's 229s. NB followed closely with an accuracy of 86.5%, with a response time of 33s, kNN had an accuracy of 86.3% and a response time of 327s, Random Forest had an accuracy of 86.3s and a response time of 219s, Decision Tree classifier is the one that had the lowest accuracy of 67.6% and a response time of 28s

From the above results it is evident that the deep learning based classifier outperforms other classifiers in predicting both the first career choice and the second followed by support vector machine.

4.3 Career Recommender System

Currently recommendation frameworks are used in areas where there is overwhelming data. This allows the clients to concentrate only on significant areas of interest. The career recommender system aims at providing a career guidance and prediction at student's level so that they comply with the guaranteed requirements. Its input consists of (1) student data (2) career selection procedure (3) cluster subjects. The prediction and career paths are the output. The system focuses on classifying student data and recommending a suitable career path.

The architecture of the career recommendation system was simulated as in Figure 4.

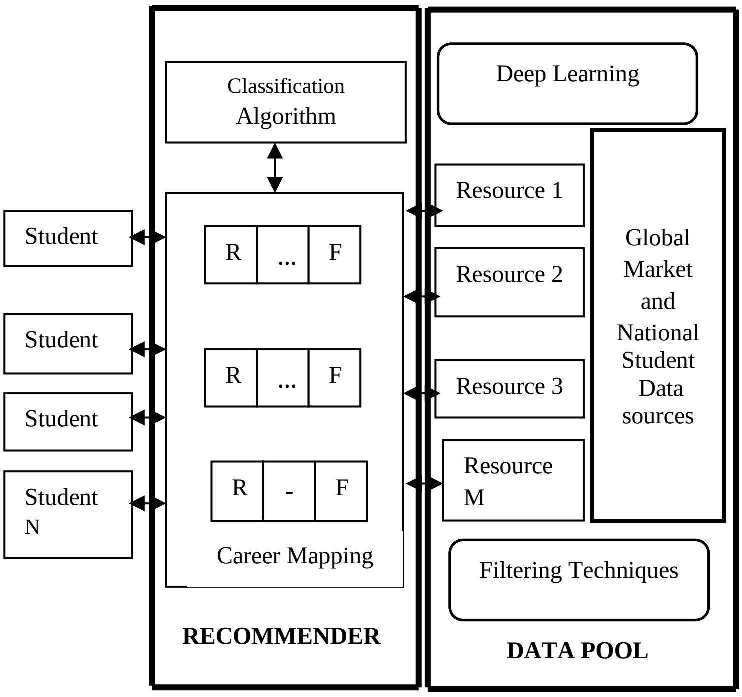

Figure 4: Architecture of Recommender Algorithm

Figure 4 presents a summary of classification algorithms for implementing Career recommender system in secondary schools. At the heart of the recommender system is the classification algorithm for classifying student data, career mapping that assigns a validated career path to a client (student or career master). The recommendation is generated and classified on priority basis i.e. highest priority path to the lowest likely path during mapping process. The F denoting the first student priority career path and R denoting the lowest student priority path which varies due different signatures of data that is unique to different students' capabilities and nature of subjects offered in the school.

The data pool define a repository of mixture of data and various techniques of accessing and machine learning techniques. The pool has unique data sources denoted as resources with a set of numbers {1, 2, .. M} as sub-servers for different school which comprises student data from different levels, global market and National student data sources which provides that on dynamics of market demands and general data as regulated in NEMIS (government planning and monitoring). Deep learning algorithms are used to enable training and learning process for accurate storage of data in the resource data set for accurate classification and career mapping. Filtering algorithm filters the relevant global data toward unified career decision making in the recommender system. Model clients i.e. students and career masters denoted as a set of student {1, 2, 3,…, N} defines the set of requests for career path services from the recommender which are handled in a hand-shake manner.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Study Summary

The purpose of pursuing this study was to evaluate the best classification algorithm for implementing the career recommendation system. To achieve the objective the researcher collected data from form four students on the three key attributes: personality, interest and academic achievement. This data was preprocessed and acted as the input in the RapidMiner data mining tool that evaluated the performance of the various classification algorithms in predicting the suitable career for a student. It was noted that deep learning and support vector machine algorithms out-performed other algorithms in this exercise as seen in tables 1 and 2.

The study further came up with an architecture of career recommender system that uses a variety classification algorithms. It was noted that recommender system for implementing Career recommendation in secondary schools has at the heart a classification algorithm for classifying student data and a career mapping algorithm that assigns a validate career path to a client (student or career master). The recommendation is generated and classified on priority basis i.e. highest priority path to the lowest likely path during mapping process. The mapping process is as represented in figure 3 having F denoting the first student priority career path and R denoting the lowest student priority path which varies due to different signatures of data that is unique to different students' capabilities and nature of subjects offered in the school. The data pool define a repository of mixture of data and various techniques of accessing and machine learning techniques. The pool has unique data sources denoted as resources with a set of numbers {1, 2, .., M} as sub-servers for different schools which comprises student data from different levels, global market and National student data sources which provides that on dynamics of market demands and general data as regulated in NEMIS (government planning and monitoring). Deep learning algorithms are used to enable training and learning process for accurate storage of data in the resource data set for accurate classification and career mapping. Filtering algorithm filters the relevant global data toward unified career decision making in the recommender system. Model clients i.e. students and career masters denoted as a set

5.2 Conclusion from the Study

The study revealed that classification algorithms can be used to predict a suitable career for a student. All the classifiers that were used gave a predictive accuracy of above 88% for the first choice though deep learning was the most accurate with 97.5%. However, since the classifiers out performed each other in various metrics, therefore using multiple classification algorithms in building the recommendation model can yield better results. The study therefore concludes that classification models comprising of multiple classifiers can be used to predict suitable careers for secondary students.

5.3 Recommendation from the Study

This study recommends that Multiple classification algorithms to be used in building the recommendation models. This is because the classification algorithms that were used to build the classification models out performed each other in various metrics, therefore using multiple classification algorithms can yield better results.

6. REFERENCES

[1] Algarni, A. (2016). Data Mining in Education. International Journal of Advanced Computer Science and Applications, 456-461.

[2] Ansari, G. A. (2017). Career Guidance through Multilevel Expert System Using Data Mining Technique. *International Journal of Information Technology and Computer Science*, 8, 22-29.

[3] Elakia, Gayathri, Aarthi, & Naren, J. (2014). Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students. *International Journal of Computer Science and Information Technologies* , Vol. 5 (3), 4649-4652.

[4] Han, J. & Kamber, M. . (2006). *Data Mining Concepts and Techniques.* 2nd ed. San Francisco: Morgan Kaufmann, pp.5-7.

[5] Mberia, A. & Midigo, R. (2018). Understanding Career Choice Dilemma in Kenya: Issues of Informed Choices and Course Availability. *Journal of Education and Practice*, Vol.9, ISSN 2222-1735 (Paper).

[6] Mohamad, S. K., & Tasir, Z. (2013.). Educational data mining: A review, . *Procedia-Social and Behavioral Sciences*, vol. 97, pp. 320–324, 2013.

[7] Mudulia, A. M. (2017). Relationship Between Career Guidance and Counselling and, Career Choice Among Secondary School Girls In Vihiga County, Kenya. (PhD Thesis). School of Education: Moi University, Kenya.

[8] Mugenda, A. & Mugenda, O. (2012). Research methods: Qualitative and Quantitative Approaches. Nairobi: Acts Press.

[9] Muhammad, S & Safawi A. R. (2017). Challenges of Applying Data Mining in Knowledge Management towards Organization. *International Journal of Academic Research in Business and Social Sciences*, Vol. 7, No. 12, 405-412.

[10] Nikita, G., Ishani, Z., Aishwarya, N., & Deepali, N. . (2017). Career Counseling using Data Mining. *International Journal of Engineering Science and Computing*, Volume 7 Issue No.4: 10271-10274.

[11] Nisbet, R., Elder, J., & Miner, G. . (2009). *Handbook of statistical analysis and data mining applications.* Amsterdam : Elsevier.

[12] Raheela, A., Saman, H. & Saba, I. H. (2017). Predicting Student Academic Performance using Data Mining Methods. *IJCSNS International Journal of Computer Science and Network Security,* , VOL.17 No.5: 187-191.

[13] Siraj, F., & Abdoulha, M. (2009). Uncovering hidden information within university's student enrolment data

using data mining. *MASAUM Journal of Computing* , 1(2), 337-342.

[14] Stephen, S. K., George, O., & Richard, R. (2016). Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions. *International Journal of Computer Applications Technology and Research*, Volume 5–Issue 11, 698-704.

[15] Suhirman, Zain, J. M., Chiroma, H., & Herawan, T. (2014). Data Mining for Education Decision Support: A review. *iJET*, Volume 9, Issue 6, pp 1-19. Retrieved from http://dx.doi.org/10.3991/ijet.v9i6.3950

[16] Wabwoba F. & Mwakondo M. (2011). Students Selection for University Course Admission at the Joint Admissions Board (Kenya) Using Trained Neural Networks. *Journal of Information Technology Education*, Volume 10, 334-345.

[17] Yadav, S., & Pal, S. . (2012). Data Mining Application in Enrollment Management: A Case Study. *International Journal of Computer Applications*, 41(5), 1-6.