Discovering Knowledge regarding Academic Profile of Students Pursuing Graduate Studies in World's Top Universities

Shibbir Ahmed[*], Abu Sayed Md. Latiful Hoque[*], Mahamudul Hasan[‡], Rahnuma Tasmin[¥], Deen Md. Abdullah[€],
Anika Tabassum[*]

Department of Computer Science and Engineering
[*]Bangladesh University of Engineering and Technology
[‡]University of Dhaka, [¥]Military Institute of Science and Technology, [€]Primeasia University
Dhaka, Bangladesh
{shibbirahmedtanvin, munna09bd, rabi.mist, annisizzler}@gmail.com

Abstract— **Each year a large number of students apply for graduate studies into well-known universities all around the world. Students trying to pursue higher study in a better institution strengthen their academic profile by obtaining better CGPA, high scores in standardized tests like GRE, TOEFL, IELTS, noteworthy publication records. Some of the applicants become successful by getting admission in desired university with financial support. Discovering knowledge from those successful applicants' entire academic profile by utilizing it for appropriate mining algorithm will be beneficial to the prospective graduate applicants. In this paper, we present a technique to discover knowledge from successful graduate applicant's entire academic records who got opportunity to study abroad with funding. We have preprocessed the dataset rigorously after gathering all relevant information from successful applicants. For Association rule mining, we have mentioned a technique to discretize the records of CGPA, Standardized test scores, research and job experience etc. Then we have applied Predictive Apriori Association rule algorithm and categorized interested association rules based on the predictive accuracy which are obtained from the dataset. Finally we extract knowledge which is very useful for the students who want to apply for graduate studies into suitable universities considering their entire academic profile.**

Keywords—Data Mining; Knowledge Discovery; Association Rule Mining; Predictive Apriori Algorithm; Higher Study Abroad.

I. Introduction

After completing undergraduate studies, many students want to pursue graduate study in a better institution outside their homeland. Those students try to build up their academic profile to make them competitive and successful candidates for graduate studies in their preferable academic institutions. Along with the undergraduate CGPA, standardized test scores such as GRE (Graduate Record Examination), TOEFL (Test of English as a Foreign Language), IELTS(International English Language Testing System) as well as research and job experience play significant role to make an applicant competent for graduate studies with financial support in a renowned university. Before applying to renowned graduate schools an applicant should be aware of the facts such as, which category of students get admission with funding? What

are the factors of applicants' academic profile really influences admission with funding? How current applicants should match their profile with previous successful applicants' profile in context of getting admission into a specific graduate university? Concepts and techniques of data mining [1, 2] are very much useful to discover such types of hidden knowledge from successful applicants' academic profile. From Bangladesh, each year noteworthy amount of students try to apply for graduate studies in renowned universities in USA, Australia, Canada, Japan, Germany, Singapore, Malaysia etc. Some of them get admission with full funding. A non-profit organization collects those students' entire academic data [3] for betterment of prospective graduate applicants of Bangladesh. Discovering knowledge from successful applicant's academic records using appropriate data mining and machine learning algorithm [4] will be crucial for forthcoming graduate applicants which is the main objective of this research study.

In this paper, we have developed a technique to discover knowledge using Predictive Association Rule Mining from academic profile of students who have already got admission into their desired academic institution with funding. Here, we have preprocessed the dataset extensively after gathering all relevant information from successful applicants. For Association rule mining, we have manipulated a technique to discretize the records of applicant's CGPA, Standardized test scores, research and job experience. After applying Predictive Apriori Association rule mining algorithm, we have categorized interested association rules based on the predictive accuracy which is a metric combining support and confidence. Finally we extract knowledge from those significant rules which may be very useful for the students who have profound desire to get graduate admission into a renowned institution with best possible financial support.

II. Background and Motivation

Pursuing graduate study in a renowned university in a renowned country is a cherished dream of many students. Those students prepare themselves from the very beginning of their undergraduate career. Getting inspired from elder ones and gathering knowledge from various sources, students build

up their academic profile since undergraduate studies. After completing undergraduate studies, students try to obtain better scores in standardized test scores such as in GRE, TOEFL, IELTS etc. Along with these, students try to enrich publication records and gathering research and job experience. But the most crucial step is to apply to appropriate graduate studies after selecting appropriate institutions matching applicants' entire academic profile. For this, appropriate knowledge is required which can be obtained from successful applicants of previous sessions who build up their academic profile and eventually get admission into desired academic institution. Here, only statistical analysis and personal observation may not be sufficient in this regard because of the nature of the data. That is why we are motivated to do research and apply data mining techniques to discover knowledge regarding academic profile of students pursuing higher study abroad with a view to assisting prospective future graduate applicants to apply for graduate studies in appropriate graduate universities.

III. RELATED WORK

Educational data mining (EDM) has been deployed to explore data originating in an educational context. There is a survey [5] which illustrates the most relevant studies carried out in this field. Mining educational data to analyze students' performance from the extracted knowledge has been presented in recent studies [6, 7, 8]. Besides these, there is also a book chapter [9] where educational data mining methods, it's ongoing trends and it's linkage with learning analytics community are presented. A latest survey [10] of educational data mining based on analysis of recent works can be visualized as well. Mining significant association rules from educational data using critical relative support approach has been proposed recently in a study [11]. There is a significant research study [12] which describes the measure of interestingness for association rules in educational data. Mining techniques of rare rules are demonstrated in a research [13] clearly. Although there is a study showing the drawbacks and solutions of applying Association Rule Mining in learning management systems [14], the mined association rules reveal various factors like student's interest, curriculum design; teaching and learning outcomes has been successfully illustrated in significant research studies [15,16].

A technique of preprocessing academic data before Mining Association Rule [17] with synthetic dataset has been proposed recently for checking the suitability of the system with the real institutional dataset. A significant survey [19] on Association Rule Mining algorithm has been proposed where it is shown that the quality of small rule sets generated by Apriori can be improved by Predictive Apriori algorithm applied on UCI Repository of machine learning datasets. Another important study [20] has compared the performance of Apriori and Predictive Apriori based on the interesting measures using Weka. A recent study [21] has done noteworthy comparison between Apriori and Predictive Apriori algorithm in which Predictive Apriori has been found better and faster than Apriori algorithm in mining association rules from a dataset containing crimes data concerning women taken from UCI repository. Knowledge discovery regarding educational dataset of a university using Association Rule mining has been presented in our recent study [18] where Apriori algorithm is applied to extract knowledge of students' personal statistics, academic performance as well as impact of courses and curriculum. However, Predictive Apriori performs better than Apriori and higher study abroad dataset requires different algorithms for preprocessing. Thus, we are encouraged to improve the technique and discover hidden knowledge regarding academic profile of students pursuing higher study abroad using Predictive Apriori Association Rule Mining algorithm.

IV. METHODOLOGIES

A. Data Analysis

Several personal and academic information of a particular student is stored in the universal table created from the information found in the successful applicant's records [3]. Those are collected for data preprocessing and data analysis.

TABLE I.   SELECTED DATA FROM UNIVERSAL DATABASE

Academic Data	Undergrad Department
	Undergrad University
	Undergraduate CGPA
	Research Area
	Research Experience
	Job Experience
	Publication Records
Standardized Test Score Data	GRE (Verbal, Quantitative, Analytical Writing, Total Score)
	TOEFL/ IELTS (Reading, Listening, Speaking, Writing, Total Score)
Personal Data	Gender

The personal and academic data as well as Standardized test scores data stated in the above Table I are considered. As we have experimented with the successful applicants' data who pursue higher study abroad, we have analyzed all records of them such as, Outgoing University, Outgoing Country, Intended Semester, Admission In (M.Sc. or PhD). We have considered these graduate admission records stated in the above Table II for knowledge discovery regarding academic profile of successful graduate applicants.

TABLE II.   SELECTED GRADUATE ADMISSION RECORDS FROM UNIVERSAL DATABASE

Graduate Admission Records	Outgoing University
	Outgoing Country
	Outgoing State in Country
	Intended Semester
	Admission In (M.Sc. or PhD).
	Funding
	Application Step

B. Preprocessing for Mining Higher Study Abroad Database

First of all, all-inclusive customized universal database is created in which records of all successful applicants from Table I and Table II are utilized. In this universal table redundant records and incomplete or null value records are omitted from the existing database for the suitability of applying mining algorithm. So, records such as name and other personal information have been omitted in the customized universal table shown in Table III. But this dataset is yet not ready for

applying association rule mining algorithm. For this we have to discretize the continuous attributes such as CGPA, test scores etc.

TABLE III. A SMALL PORTION OF CUSTOMIZED UNIVERSAL DATABASE

Gen-der	Outgoing Country	Outgoing University	Depar tmnt	Research Interest	GRE Score	TOE FL/IE LTS Score	Pub licat ions	Fun ding	
Male	USA	UC Berkley	EEE	Signal Processing	330	115	10	TA	…
Fe-male	USA	UIUC	CSE	Data Mining	315	90	2	RA	
…	…	…	…	…	…	…	…	…	

The customized universal database of Table III has been transformed into an equivalent transformation table by transforming the continuous valued attribute as discrete valued attribute representing some knowledge for the suitability of implementing Predictive Apriori algorithm of Association Rule Mining. As for example, CGPA, GRE, TOEFL, IELTS score are continuous attributes and it has been transformed into five classifications as Very High, High, Medium, Low and Very Low for better analysis. One algorithm has been used for discretizing all continuous scores of GRE, TOEFL and IELTS. Another algorithm has been used for transforming undergraduate CGPA into those five classifications. For transforming the CGPA of latest academic level, Algorithm 1 has been developed to populate the transformed table in such a way that there is no continuous CGPA value in an entry.

Algoithm 1: CGPA_Discretization( )

Input: all acquired *CGPA* of a student in undergraduate level of study in the universal table
Output: *transformed_ CGPA* for the Transformation Table
for *i*=1 to | *Student$_{list}$* |
   if (*CGPA>= 3.75 && CGPA<=4.00)*
     transformed_CGPA= 'VeryHigh'
   else if (*CGPA>= 3.50 && CGPA<3.75)*
     transformed_CGPA = 'High'
   else if (*CGPA>= 3.25 && CGPA<3.50)*
     transformed_CGPA = 'Medium'
   else if (*CGPA>= 3.00 && CGPA<3.25)*
     transformed_CGPA = 'Low'
   else if (*CGPA<3.00)*
     transformed_CGPA = 'VeryLow'
end for

Similarly the scores of all standardized test scores in customized universal table are also transformed by an algorithm named as 'Test Score Discretization' in Algorithm 2. As the real data set contains test scores of GRE, TOEFL, IELTS with various range we have considered another variable discretization format table and transformed the continuous value of various test scores to these five classified definitions.

To construct the entire transformed table as given in Table VII, we have used the customized universal table and stated discretized format of Table IV, Table V and Table VI.

TABLE VII. TRANSFORMED TABLE FROM CUSTOMIZED UNIVERSAL TABLE

Gender	Outgoing Country	Outgoing University	Department	ResearchArea	CGPA	GRE_Score	TOEFL/ IELTS_Score	Funding	
Male	USA	UC Berkley	EEE	Signal Processing	VeryHigh	VeryHigh	VeryHigh	FullWaiver_TA	……
Female	USA	UIUC	CSE	Data Mining	Medium	High	Medium	PartialWaiver_RA	
….	….	….	….	….	….	….	….	….	

Algoithm 2: TestScore_Discretization( )

Input: *scores of GRE(verbal, quant, awa), TOEFL, IELTS (reading,listening,speaking,writing)* from Universal Table of *Student$_{list}$*
Output: discrete *level* of *scores* for the Transformation Table
for *i*=1 to | *Student$_{list}$* |
   if (*score>=90%*)
     level = "VeryHigh"
   else if (*score>=80% && score<90%*)
     level = "High"
   else if (*score>=65% && score<80%*)
     level = "Medium"
   else if(*score>=50% && marks<65%*)
     level = "low"
   else if(*score<50%*)
     level = "VeryLow"
end for

As there are standardized test scores of GRE, TOEFL and IELTS with different range of scores, it requires different discretization format tables for all these different types of scores based on real dataset. For example, Discretization format table for GRE scores (in Table IV), for TOEFL (in Table V), and for IELTS (in Table VI) are illustrated below.

TABLE IV. DISCRETIZATION FORMAT TABLE FOR GRE SCORES

Classified Name	Range of Scores (S)			
	Verbal	Quant	AWA	Total
VeryHigh	165≤S<170	165≤S<170	4.5<S≤6.0	320≤S≤340
High	160≤S<165	160≤S<165	3.5<S≤4.5	310≤S≤320
Medium	150≤S<160	150≤S<160	3≤S≤3.5	300≤S<310
VeryLow	140≤S<150	140≤S<150	2≤S≤2.5	290≤S<300
Low	130≤S<140	130≤S<140	S<2	260≤S<290

TABLE V. DISCRETIZATION FORMAT TABLE FOR TOEFL SCORES

Classified Name	Range of Scores (S)	
	Reading/Listening/Speaking/Writing	Total
VeryHigh	27≤ S ≤30	108≤S≤120
High	24≤ S < 27	95≤S<108
Medium	18 ≤ S < 24	80≤S<95
VeryLow	10 ≤ S < 18	60≤S<80
Low	*S <10*	S<60

TABLE VI. DISCRETIZATION FORMAT TABLE FOR IELTS SCORES

Classified Name	Range of Scores (S)	
	Reading/Listening/Speaking/Writing	Total
VeryHigh	S > 7.5	S > 7.5
High	7.0 ≤ S ≤ 7.5	7.0 ≤ S ≤ 7.5
Medium	6.0 ≤ S ≤ 6.5	6.0 ≤ S ≤ 6.5
VeryLow	5 ≤ S ≤ 5.5	5 ≤ S ≤ 5.5
Low	S <5	S <5

We have considered Job and research experience in terms of years and publication records as five classified values (0: VeryLow, 0-2:Low, 3-4:Medium, 5-6:High; >6:very High) based on numeric value given in the real dataset of academic profile of successful graduate applicants. All other attributes in the customized universal are discrete and suitable for Association Rule mining.

V. EXPERIMENTAL PLATFORM

In this experiment, data up to the last three years including different sessions such as 'Fall', 'Spring' are considered. There are more than a thousand records of successful applicants who are now pursuing higher study abroad with financial support. We have categorized relevant academic and standardized test records as well as job and research experience and built into a customized universal table structure of about four hundreds successful applicants ignoring missing values as well as null values. Finally we transformed it into a transformed table structure which contains only discretized attributes for applying Predictive Apriori Association Rule mining algorithm.

After preprocessing step, a customized transformed table of about 392 students of different universities of Bangladesh has been obtained who have already got opportunity to study MSc or PhD in the renowned universities abroad. We have manipulated the transformation table containing all continuous data like CGPA, test scores and transformed into five discrete values. Then, we have used Weka Explorer [22], a popular suite of machine learning software written in Java. We have applied Predictive Apriori algorithm to the customized transformation table with predefined minimum support and confidence using Weka to generate interesting Association Rules. Choosing predictive accuracy precisely and selecting important association rules from huge number of generated rules are very important in this regard.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. Influence of Gender

The influence of gender has been found in the overall academic profile of successful applicants. We have selected just few rules related to gender of successful applicants with higher predictive accuracy. There are multiple factors that affect admission with funding and academic records. The result of Fig. 1 points out that the male students have very high predictive accuracy of getting funding as RA with 1 year research experience. Another rule indicates that getting PhD admission with full funding as TA among female students is better than the male students.
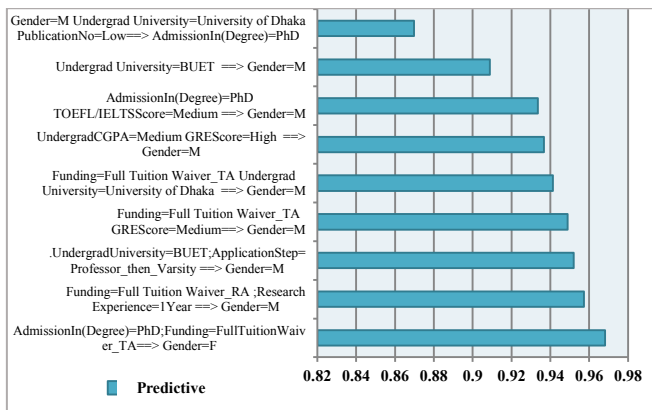

Fig. 1. Influence of Gender

B. Influence of Undergraduate Department & University

In Fig. 2, there are several interesting association rules with high predictive accuracy. One Rule indicates that if admission opportunity is obtained by a Civil Engineer, then he usually achieve funding in previous years. Again another implies that if a BUET graduate tries to apply for PhD with funding, he or she should apply to a university after contacting with professor.
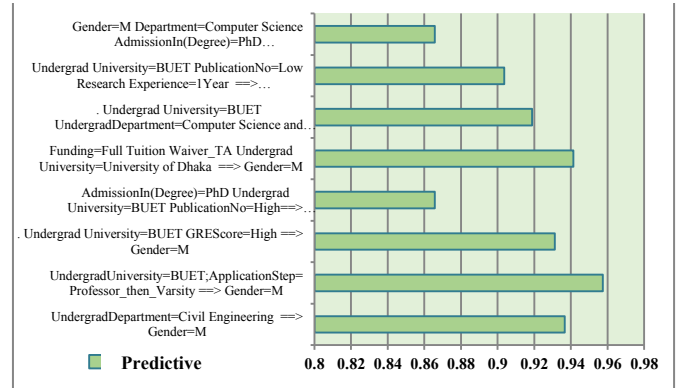

Fig. 2. Influence of Undergraduate University & Department

C. Influence of Research Area

The analyzed Association Rules show that research area influences funding as RA. In rule number 6, we find that if anyone has research interest with Data Mining he or she may get funding with RA as the rule has higher predictive accuracy than research area of bioinformatics. Again according to rule no. 3, if anyone does not have fixed research interest, he or she should focus on getting high GRE scores for funding opportunity according to the rule stated in Fig. 3.
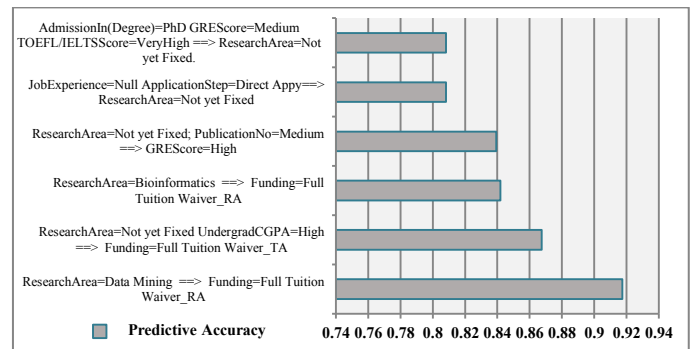

Fig. 3. Influence of Research Area

D. Influence of Undergraduate CGPA

We have selected few interesting association rules based on higher predictive accuracy presented in the Fig. 4. For example a rule indicates that if any student has very high CGPA and high GRE Score, he or she should contact professor before applying to a university. Again another significant rule indicates that the number of male students with undergrad CGPA high is relatively large in the last three years dataset of successful applicants.
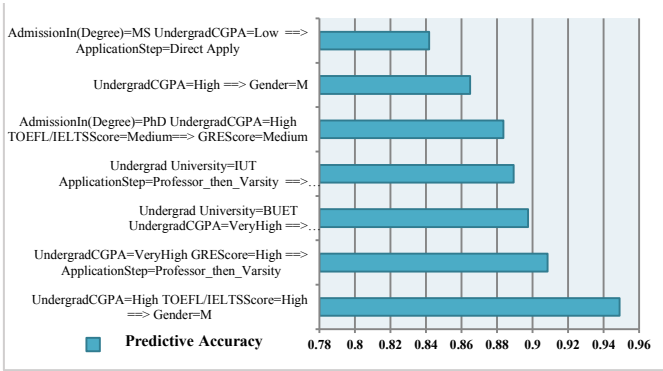
Fig. 4. Influence of Undergraduate CGPA

Fig. 6. Influence of TOEFL/IELTS Score

E. Influence of GRE Score

The students who get medium range of GRE scores generally gets high score in quant but low in verbal. By analyzing the rules illustrated in the Fig. 5, we have observed such kind of fact. It is discovered that with a higher predictive accuracy that students with high GRE scores should direct apply to university and funding opportunity will be relatively higher in this kind of cases. Again, if a student applies for PhD and gets full funding as RA; he or she usually gets very high GRE score as a significant rule of below figure indicates.
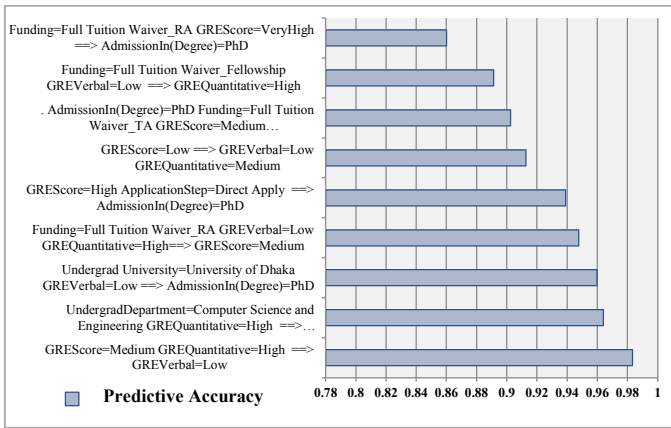
Fig. 5. Influence of GRE Score

F. Influence of TOEFL/IELTS Score

Several significant rules can be selected to indicate the influence of TOEFL/IELTS score based on higher predictive accuracy. One rule in the Fig. 6 indicates that if a student with low CGPA gets funding then his or her TOEFL score is very high. Again, another significant rule indicates that if a student with low CGPA directly applies to a university abroad, he or she gets full funding because of high TOEFL/IELTS speaking scores. Another important rule indicates that if a very low CGPA holder applies after getting high speaking and listening score in TOEFL/IELTS; he or she gets admission in PhD with funding.
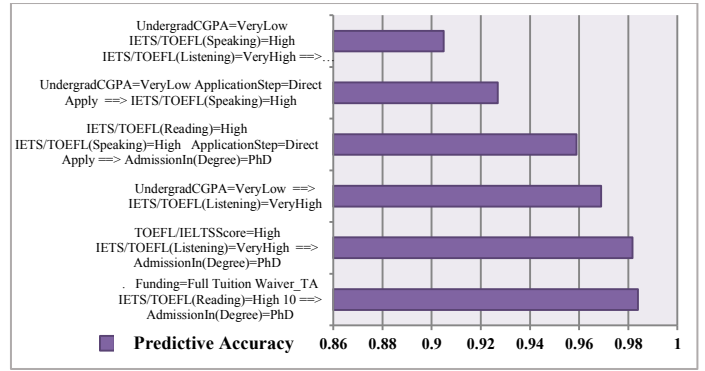
G. Influence of Research Experience & Publication Record

After analyzing the generated Association Rules (in Fig. 7) we observed various impacts of research experience and publication records of a successful applicant. According to the fifth rule if an applicant has research experience more than 2 to 4 years and applies for PhD admission, the probability of getting full tuition waiver with RA is much higher. Again another rule implies that, if an applicant has very high publication records and he or she applies after contacting with professor, he or she usually gets PhD admission with funding.
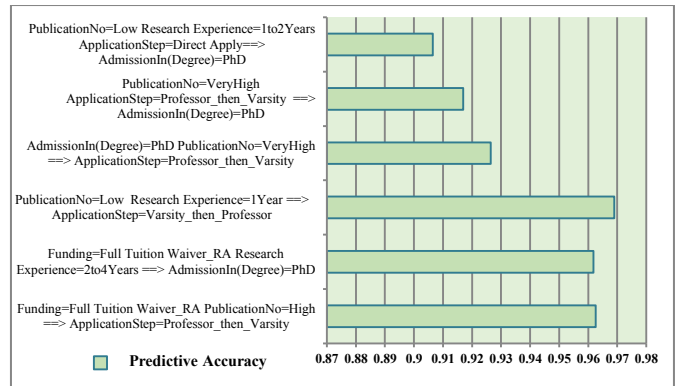

Fig. 7. Influence of Research Experience & Publication Record

H. Influence of Job Experience

From the generated association rules based on predictive accuracy we have analyzed influence of job experience of the applicant and stated few significant rules in Fig. 8. For example, according to the last rule in the Fig. 8, if an applicant has job experience of more than 4 years and contacts with professor before applying; he or she may probably get admission in PhD program. Again, second rule implies that if a fresher applies for PhD program with no job experience and he or she directly applies to university without contacting with any professor, the probability of getting full tuition waiver with RA is higher according to the generated predictive accuracy 0.90342.
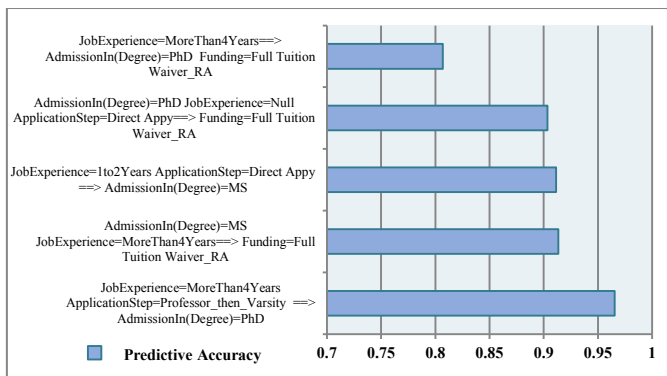
Fig. 8. Influence of Job Experience

VII. CONCLUSIONS

Discovering knowledge from entire academic profile of successful graduate applicants and applying it properly for providing or achieving guidelines will be very much beneficial for the prospective graduate applicants. Here, we have considered the data of three years which are voluntarily given by the successful applicants, but there many other students who get admission with funding yet to record in the list. We can apply the same technique to extract knowledge from all the students pursuing higher study abroad in last decade. Again, we have considered the students of Bangladesh going abroad for graduate studies. We can modify the technique so that it can be applicable to the analysis of applicants of other countries like Indonesia, Malaysia, China etc. In this research, we have preprocessed the dataset efficiently after gathering all relevant information from successful applicants before applying the data mining algorithm. For Association rule mining, we have manipulated a technique to discretize the records of applicant's continuous attributes like CGPA, Standardized test scores of GRE, TOEFL and IELTS, research and job experience etc. Then we have applied Predictive Apriori Association rule mining algorithm to generate association rules. We have categorized interested association rules based on higher predictive accuracy. Finally we extract knowledge from those significant rules which may be very helpful for the students who have profound desire to obtain graduate admission into a renowned institution abroad with financial support.

REFERENCES

[1] J. Han , M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques", 3rd ed. San Francisco:Morgan Kaufmann Publishers, 2011.

[2] Aggarwal, Charu C. "*Data mining: the textbook*". Springer, 2015.

[3] (2015, Oct.) All bangladeshis abroad Universal Database [Online]. Available:http://www.higherstudyabroad.com/database/all-bangladeshis-abroad/

[4] Charu C. Aggarwal, "Data Mining: The Textbook", 1st ed., Springer International Publishing Switzerland, 2015.

[5] Romero, C., & Ventura, S. "Educational data mining: a review of the state of the art". IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618, 2010.

[6] Baradwaj, B. K., & Pal, S. "Mining educational data to analyze students' performance". arXiv preprint arXiv:1201.3417, 2012.

[7] El-Halees, A. "Mining students data to analyze e-Learning behavior: A Case Study". Department of Computer Science, Islamic University of Gaza PO Box, 108, 2009.

[8] Romero, C., Ventura, S., & García, E. "Data mining in course management systems: Moodle case study and tutorial". Computers & Education, 51(1), 368-384, 2008.

[9] Baker, R. S., & Inventado, P. S. "Educational data mining and learning analytics". In Learning analytics, pp. 61-75, Springer New York, 2014.

[10] Peña-Ayala, A. "Educational data mining: A survey and a data mining-based analysis of recent works". Expert systems with applications, 41(4), 1432-1462, 2014.

[11] Abdullah, Z., Herawan, T., Ahmad, N., & Deris, M. M. "Mining significant association rules from educational data using critical relative support approach". Procedia-Social and Behavioral Sciences, 28, 97-101, 2011.

[12] Merceron, A., & Yacef, K. "Interestingness measures for association rules in educational data". In Educational Data Mining, 2008.

[13] Romero, C., Romero, J. R., Luna, J. M., & Ventura, S. "Mining rare association rules from e-learning data". In Educational Data Mining, 2010.

[14] E. Garcia, C. Romero, S. Ventura, T. Calders, "Drawbacks and solutions of applying association rule mining in learning management systems," In Proceedings of the International Workshop on Applying Data in e-Learning, 2007.

[15] Kumar, V., & Chadha, A. "Mining association rules in student's assessment data". International Journal of Computer Science Issues, 9(5), 211-216, 2012.

[16] Parack, S., Zahid, Z., & Merchant, F. "Application of data mining in educational databases for predicting academic trends and patterns". In Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on (pp. 1-4). IEEE, 2012.

[17] A.S.M. L. Hoque, R. Paul, S. Ahmed, "Preprocessing of Academic Data for Mining Association Rule," In Proceedings of the Workshop on Advances in Data Management: Applications and Algorithms(WADM), 2013.

[18] Ahmed, S., Paul, R., & Hoque, A. S. M. L. "Knowledge discovery from academic data using Association Rule Mining". In Computer and Information Technology (ICCIT), 2014 17th International Conference on (pp. 314-319). IEEE, 2014.

[19] Stefan Mutter, Mark Hall and Eibe Frank. "Using Classification to Evaluate the Output of Confidence-Based Association Rule Mining". In the Proceedings of 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. DOI: 10.1007/978-3-540-30549-1_47.

[20] Sunita B. Aher and Lobo L.M.R.J. "A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning". International Journal of Computer Applications (0975 – 8887), Volume 39– No.1, February 2012.

[21] Divya Bansal and Lekha Bhambhu. "Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women". International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 9, September 2013.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update." *SIGKDD Explorations*, 11(1), 2009.