Predicting Academician Publication Performance using Decision Tree.

Mohd Zakree Ahmad Nazri, Rohayu Abd Ghani, Salwani Abdullah, Mas Ayu, R Nor Samsiah.

Abstract:This research focuses on predicting academician performance in terms of publication rate and investigate the factors that affect academicians' achievement. This study investigates how scientific publication rate by individual is influenced by factors such as gender, age, number of research grant and academic position of the researchers. Having a decision rules, university leaders can understand upcoming trends with respect to leadership requirements and academicians needs. It is also helping university managements understand challenges and therefore can deploy the right strategies for human resource management interventions. This paper describes the development of the predictive model using a data mining technique. Previous studies have shown that there are many important variables when analysing academicians' productivity at the individual level. What is unusual with our approach is that this study is using Decision Tree to identify the patterns for predicting next year's performance. Decision Tree, C4.5, J48 and PART is a common predictive method for prediction as there are other methods that are better suited for predictive analytics such as regression or metaheuristic algorithms. However, with finding knowledge among the attributes obtained from the university's databases, we can predict the performance of an academician staff. To find strong and valid rules, different measures like min Interest, lift, leverage and conviction are considered. The study, involving almost 3000 university lecturers, shows a number of interesting patterns that can be used for predicting individual performance.

Index Terms*: *Higher Education Institution, Predictive Analytics, Tree to Rule Induction.

I. INTRODUCTION

In 2006, Malaysia Ministry of Higher Education (MoHE) identified five universities in Malaysia to become the premiere Malaysian Research University (MRU). MoHE developed a comprehensive system to assess the performance of the MRU called the Malaysian Research Assessment Instrument (MyRA). MRU and all other institution of higher learning were mandated to participate in the assessment exercise. Research grant especially international grant, principal investigator or research project leader roles and high impact publications are among the main performance indicators in MyRA. Percentage of journal in first quartile (Q1) journals or high impact publications level have been among the most critical criteria and was rated extensively higher than other type of publication. Total citations also receive high score not just for the sake of MyRA assessments but also for yearly performance appraisal and promotion of an academician. Growth of publication outputs such as total number of papers in indexed journals and number of citations can be seen as the most important criteria in academic promotions, job retention, and professional development of a faculty members.

Annual performance appraisal and promotion decisions in almost all Malaysian universities are derived from the tripartite model which is, an academic staff's performance is evaluated based on three aspects, that is, (1) teaching, (2) publication and (3) service to the university and community (committee duties, administrative works, consultation and etc.) (Dhillon, Ibrahim, &Selamat, 2015). Having publication as one of the important key performance indicator (kpi) for academia has led to increased number of publications, especially in indexed journals by SCOPUS and Web of Science (WoS). The increased number of publication in indexed and high impact journals is the results of implementing effective performance management in human resource (HR) - what gets appraised, gets improved (Turk &Killumets, 2014).

An effective performance management has become an integral aspect of a university's human resource management especially under the MyRA framework. With the emergence of Fourth Industrial Revolution (4IR), it is crucial for MRU such as UniversitiKebangsaan Malaysia (UKM) to incorporate data and analytics into its core strategic vision in developing and innovating the right human resource management processes in building talent. UKM has been actively pursue effective HR management using business intelligence technologies to study the past, understand what is happening and to 'know' the future. Dunphy& Hackman (1988) believes that appraisal should be a proactive intervention addressing future issues, supporting rather than subverting corporate strategy. Thus, a HR predictive model is needed for a proactive intervention. Predictive model can be used to support HR in planning training and leadership development that facilitate training need analysis to transfer the right skills and knowledge to the right group of staff at the right time. A good model will allow HR Development Department to tailor the training plan to the level of function or generic based on the need of an individual or small group of staff so that they can learn and apply the appropriate skills together.

Revised Manuscript Received on June 7, 2019.

Mohd Zakree Ahmad Nazri, Mohd Center for Professional and Leadership Development, UniversitiKebangsaan Malaysia, zakree@ukm.edu.my

RohayuAbdGhani,Mohd Center for Professional and Leadership Development, UniversitiKebangsaan Malaysia.

Salwani Abdullah,MohdCenter for Artificial Intelligence, Faculty of Information Science & Technology, UniversitiKebangsaan Malaysia.

Mas Ayu,MohdCenter for Artificial Intelligence, Faculty of Information Science & Technology, UniversitiKebangsaan Malaysia.

Nor Samsiah,MohdCenter for Artificial Intelligence, Faculty of Information Science & Technology, UniversitiKebangsaan Malaysia.

Retrieval Number: B10260782S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Predictive modelling is a term referring to the process of knowledge extraction from historical records, process and validate a model that can be used for predicting employee performance. One of the common approach in predictive modelling is to use a data mining technique that attempts to answer the question “what will happen in the future?”. In the context of university HR predictive analytics, ‘predictive’ means identifying the quantitative factors that can be used to predict performance or risk. In research grant application assessment for example, it may be the case that those applicants who have certain demographic characteristics combined with historical data of past performance behaviour are highly likely to fail in delivering what was required (e.g. publications, intellectual property, product commercialization, etc.), in line with the research grant objectives.

The main question for this research is to answer question - Is it possible to predict academician performance in publication? This study aims to fulfil the HR requirement for a predictive model by developing a predictive analytic framework based on decision tree and rule induction to generate rules relating personnel information with publication performance. The model can be used and applied in many ways such as a guidance for HR to produce specific recruitment and HR management strategies. The second objective of this study is to investigate factors associated with scholarly publication productivity among academic staff. The paper is divided into several sections as follows: Section 2 reviews the literature, followed by an outline of existing factors and determinants found in literature. Research methodology will be discussed in Section 3 where data collection and data mining techniques are discussed. Then, Section 4 discuss the data analysis and results. Lastly, Section 5 summarizes the conclusions of this study with recommendations and direction for further research suggested.

II. LITERATURE REVIEW

A+The Data Warehousing Institute (TDWI) define business intelligence (BI) as the tools, technologies, and processes required to turn data into information and information into knowledge and plans that optimize business actions (Eckerson, 2007). A business intelligence tool with a HR analytical dashboard is required to enable access to and analysis of HR information to improve and optimize decisions and organizational performance along the MyRA framework. Among the first model that UKM HR Department required was the predictive model for research productivity, especially regarding the publication where HR and the Deaneries are concerned with improving individual performance through interventions.

The logic behind the demand for publication predictive model by the UKM HR can be learnt from the Oakland Athletics’ unconventional approach in building a successful baseball team (MacLennan, 2005). The Oakland Athletics’ general manager, Billy Beane hired a statistician and applied predictive analytics to win more games with less money (Waller & Fawcett, 2013). Billy Bean used predictive analytics to predict the potential success or failure of baseball players by using historical data. The same principle can be applied by universities to achieve the objective of a research university and maintain advantages in era of 4IR.

A. Analytics in Research Productivity

There has been a considerable number of investigation of factors and determinants between academic staffs in relation to publication output. There are strong claims that low efficacy leads to low motivation and insufficient effort, leading to lower publication output. Numerous studies conducted by researchers such as Maltby et al. (1995), Hemmings& Kay (2010), Epstein & Fischer (2017) or Reyes-Cruz &Perales-Escudero (2016) have shown that academics or scientists with low levels of self-efficacy tend to produce low research output. Bandura (1977) define self-efficacy as an individual's belief in his or her capacity to execute behaviors necessary to produce specific performance attainments. Self-efficacy reflects confidence in the ability to exert control over one's own motivation, behavior, and social environment (Bandura, 1977).

The knowledge discovered and represented as determinants or factors can be modelled into a predictive model. A publication output predictive model could support human resource managers in deciding apt vital enhancement trainings or other type of interventions such as counselling for employees to efficiently respond to performance evaluations and expectation (Kirimi&Moturi, 2016). Other than training need analysis, predictive analytics have been used to improve personnel selection and enhance human capital (Chien& Chen, 2008). Having a predictive model that can forecast the performance of a scientist can be an effective personnel selection mechanism to find talents who are the most suitable to the university.

Bland et al (2005) conducted a study to tests the ability of the Bland et al. (2002) model to predict a medical faculty and department research productivity. Bland et al (2005) used data from a University of Minnesota Medical School and multiple logistic regression technique to build a model of relationship between faculty a faculty member's level of research productivity and individual, institutional, and leadership qualities. They concluded their research that their model is able to predict the productivity of individuals and departments based on three broad groupings which are individual, institutional characteristics and effective leadership. The Bland et al. (2002) model uses knowledge extracted from two regressions to provide a predictive model that can guide HR or faculty leaders wishing to increase faculty research productivity.

Mueller et al. (2016) developed and tested a model for the prediction of countries’ research output in international evaluation journals by predictors from the research, economic, and social/political system. The developed model provided accurate predictions of countries’ research output where research productivity in the social sciences had the strongest effect, followed by economic prosperity, control of corruption, and age of evaluation society. Mueller et al. (2016) used a cross-sectional study which is a type of observational study that analyzes data at a specific point in time.

Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

predict future outcomes and trends (Beals, 2018). It uses a number of techniques, including data mining, statistical modelling and machine learning to help analysts make future business forecasts. Data mining refers to the non-trivial process of identifying novel, potentially useful and valid patterns in data (Piatetsky-Shapiro, Fayyad, & Smith, 1996).

Hemmings& Kay (2010) in their paper reviewed a number of factor and one of the important factor associated with qualifications and frequently linked to research output

is academic rank. Blackburn & Lawrence (1995) and Smeby& Try (2005) have generally reported a positive correlation between academic rank and publication output.

A number of reasons have been discussed by Hemmings& Kay (2010) given the positive correlation and one of the reason is professors have greater access to the research community networks and resources, such as doctoral students.

There is a broad set of data mining application in many domains such as military, healthcare, education and including human resources (HR) management and development. In recent years, a broad range of data mining methods have been applied in HR analytics for the four major categories of HRM Domains which are Staffing, Development, Performance Management and Compensation.

Jantan et al. (2009) developed HR system architecture for talent forecasting by learning from historical HR records. This study suggests the potential of data mining to predict performance of human resource of a public university. In their studies, they proposed a number of data mining techniques for performance prediction in HR applications:

1. Artificial Neural Network for or optimization, function approximation, patternclassification and modelling.
2. Decision Tree or Support Vector Machine for classification and prediction tasks.
3. Rough Set Theory for diagnostic analytics in understanding rules in the presence of uncertainty.
4. Fuzzy Clustering for constructing relations among data and to transform relations into knowledge.

III. METHODOLOGY

A. Research Design

The predictive modelling in this paper uses the Cross Industry Standard Process for Data Mining (CRISP-DM) model. CRISP-DM is chosen as the methodology because it is the most widely-used analytic process standard (Brown, 2015). Shearer (2000) structured the CRISP-DM process model into six major phases as shown in Fig. 1:

1. Business Understanding: Get a clear understanding of the problem you’re out to solve, how it impacts your organization, and your goals for addressing it
2. Data Understanding: Inspect, describe and evaluate the available data.
3. Data Preparation: Take data from the state it’s in to the state needed for analysis.
4. Modeling: Use mathematical techniques to make models (equations or other logic) you can use to support business decisions.
5. Evaluation: Figure out whether your models are any good.
6. Deployment: Integrate models into everyday business.

However, in this research, the model is still being studied and therefore will not be deployed until we have rebuild the model with other state-of-the-art methods.

A series of experiments were conducted to test the model using attributes extracted from the employee appraisal form.

B. The Dataset

Primary data were rigorously used in this study. The extract, transform and load (ETL) process was involved in extracting data from various databases. The UKM Centre for Information Technology (CIT) transformed the data to fit the modelling needs, then loading it into the end target data warehouse and database. Before the loading take place, the data steward followed the process of Privacy Preserving Data Mining (PPDM) by removing all sensitive private information for each record. Attribute such as name, identification number and address and other confidential data was sanitized or suppressed before it is shared with researchers.

The data gathered addressed the three core variables, namely, personal and publicatio. The dataset are collected from the university’s information systems in collaboration with the CIT and Human Resource Department (HRD). These data were extracted from various databases and stored in a data warehouse before it was pre-processed. The data set contains information from the following system’s databases:

1. Student Information Systems which details about the PhD and Master students’ supervision record.
2. Human Resource Management Systems which contain profiles and details about academic staff including the social and geographical aspects of the individual. Members of the faculty are given access to generate and publish CVs through this system.
3. Publication Management Systems that gives access to academic staff to manage research publications. Measures of the publication types includes types of publication produced throughout their service years in the institution.
4. Leave Management Systems that manages staff’s request for time away from duties and supervisors use to grant or deny leave based on government policies.
5. Research & Innovation Management Systems that provide access to the university’s research administrators and researchers to manage research grants and research publications. Users also are using this system to recruit student as research assistant

Retrieval Number: B10260782S19/19©BEIESP
Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

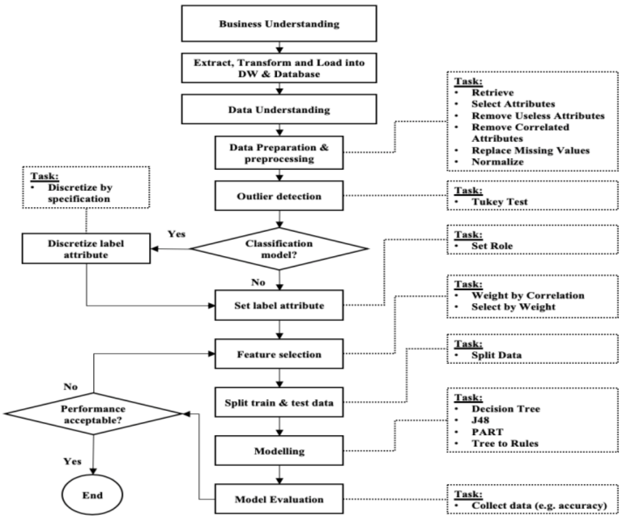

Fig. 1. The data pre-processing and predictive modelling process

The data set considers 3,306 individuals organized into two clusters. The total number of records are 27,199. The first clusters contain 1,226 academic staff (37%) who works in the discipline of humanities, arts and social science (HASS). The second cluster contains 2,080 (63%) academic staff who works in STEM (science, technology, engineering and mathematics). Each person are represented with 83 characteristics (attributes) obtained from the information systems described above. The predictive attributes were extracted from CIT's databases were used to develop a classifier. Table 1 displays a sample of the attributes from the 83 attributes.

Table 1: Sample of attributes

1) Age	2) Designation	3) No. of Research Grant (Head)
4) Gender	5) Performance Appraisal Score	6) No. of Research Grant (Member)
7) Marital Status	8) Working Status	9) Amount of Research Grant Spent
10) Depart ment	11) Administrative Post	12) No. Phd Student (Supervisor)
13) Faculty/ Institute	14) Invitation as Keynote Speaker	15) No. of Journal Article (Indexed)

Age Designation No. of Research Grant (Head)
GenderPerformance Appraisal Score No. of Research Grant (Member)
Marital Status Working Status Amount of Research Grant Spent
Department Administrative Post No.Phd Student (Supervisor)
Faculty/Institute Invitation as Keynote SpeakerNo. of Journal Article (Indexed)

C. Data Pre-processing

For better manipulation of data, the pre-processing step is paramount. As shown in Figure 1, a number activities were explored in this stage as followed:

1. First Feature Selection: This pre-processing task removes four kinds of useless attributes which are a.) Nominal attributes where the most frequent value is contained in more than the specified ratio of all examples; b.) Numerical attributes where the Standard Deviation is less than or equal to a given deviation threshold; c.) Such nominal attributes where the value of all examples is unique and d.) any attributes that satisfy the criteria given by the analyst such as correlated attributes. Correlated attributes are removed from the dataset Correlation is a statistical technique that can show whether and how strongly pairs of attributes are related. Correlated attributes are usually removed because they are similar in behavior and will have similar impact in prediction. Tukey Test is applied in this research in order to remove useless attribute by finding out which attributes are significantly different or outlier, and which are not. Outlier attributes were removed from the dataset.
2. Replace Missing Values: For each missing values are replaced by the average value of that attribute.
3. Discretize: This process discretizes the selected numerical attributes into user-specified classes. The selected numerical attributes will be changed to nominal attributes.
4. Normalize Data: Data normalization is important to scale values so they fit in a specific range as the dataset contain attributes of different units and scales. For example, appraisal scores are in percentage while total number of publication is integer.
5. Second Feature Selection by Weight: After discretization, the weight of attributes with respect to the label attribute by using correlation are calculated. The higher the weight of an attribute, the more relevant it is considered.

D. Build Model Using Machine Learning Techniques

In the context of this research, predictive modelling is a technique used to identify employee or prospects who, given their demographic characteristics, past performance appraisal and productivity, are "Meet Expectation" (ME) or "Need Improvement" (NI) in the near future. This research has found that some faculty measures the efficiency of an academic staff by their average-number of publication per-year over a longer period (e.g 3 years). Therefore it is not fair to predict and classify academician into 'underperformance' or 'highly-productive' based on yearly productivity as we seconded Starovoytova (2017) findings that there is a lack of consistent or standard in national-guidance, or institutional-policy, on how-many-publications, an-average-faculty member should-produce, per-year, to-provide a-reliable-benchmark, for-comparison. The classification process was carried out with four different Data Mining algorithms, Decision Tree (C4.5), PART, Tree to Rules and J48 to identify the best and most suitable predictive model

1. C4.5 algorithm: The C4.5 technique is one of the decision tree families that can produce both decision tree and rule-sets (Jantan, Hamdan, & Othman, 2010).
2. J48 algorithm: J48 is the improved version of the C4.5 (Jantan et al., 2010).
3. algorithmPART algorithm generates a decision list using the divide-to-conquer technique proposed by Frank & Witten (1998). At each iteration, this algorithm will place the best attribute within a rule. It builds a partial C4.5 decision tree in each

iteration and makes the "best" leaf into a rule.

4. Decision Tree used in this study is Rapidminer's version of decision tree and it is not similar to C4.5 or J48. According to Rapidminer's description of the algorithm , it is a tree like collection of nodes intended to create a decision on values affiliation to the class ("Decision Tree," n.d.). Each node represents a splitting rule for one specific attribute. For classification this rule separates values belonging to different classes in order to reduce the error. The building of new nodes is repeated until the stopping criteria are met. A prediction for the class label Attribute is determined depending on the majority of records which reached this leaf during generation.

E. Evaluation

Two metrics were used to analyze the performance of the classification of the models from the constructed attribute subsets [21, 22]:

1. Accuracy (Acc): presents the proportion of correct predictions. The calculation is derived from the confusion matrix (Table 1) and can be represented as the equation (1).
2. Spearman Rho The rank correlation between the actual and predicted labels, using Spearman's rho. Spearman's rho is a measure of the linear relationship between two variables. The two variables in this case are label attribute and prediction attribute.

IV. RESULTS AND FINDINGS

In this study as shown in Figure 1, we implement feature selection by calculating the relevance of the attributes by computing the value of correlation for each 83 attributes in the dataset. This weighting scheme is based upon correlation and it returns the absolute or squared value of correlation as attribute weight. Later, we select attributes with weights larger than the threshold (0.2) that lead to better performance and less execution time. Table 2 shows the result of implementing the feature selection by weight. The dataset has been divided into two separate clusters, STEM and HASS. STEM has 15 selected attributes while HASS is left with 9 of attributes. The selected features will be used to classify the datasets with four different classifiers to test the performance of these features.

Table 2: Two datasets and number of attributes

1) Dataset (cluster)	2) STEM	3) HASS	4) Original Dataset
5) No of Records	6) 17,345	7) 9,854	8) 27,199
9) No. of Selected Attributes used for modelling	10) 15	11) 9	12) 83

The following step refers to the previous sets the algorithms of induction of classification rules (C4.5, PART,J-48 & C4.5). According to Table 3, which presents the evaluation of the performance of the classifiers, the sets maintained an accuracy of approximately 75% and they did not present any significant performance gain when compared to each other.

Table 3: Performance Evaluation

Data Mining Technique	**Accuracy**
Decision Tree	70.3%
PART	75%
J-48	75.3%
C4.5	70.2%

Using the predictive model produced by Decision Tree, 70.31% of all predictions done by this model are correct. However the outcome is most likely Need Improvement (NI), but the model is not very confident. In fact, the confidence for this decision is only 70.24%. The attribute value of Number of Research Grant as Member (Active + Completed Project), Number of Leaves Taken Yearly, Number of Masters Student (Supervised) and Number of Research Grant as the Head of Project (Active + Completed Project) does not support this decision. When the model predict Need Improvement, it covers 93.86% of those cases. And it is correct with 69.09% of all predictions for class Need Improvement (NI).

C4.5 suffers the same problem as Decision Tree where 70.19% of all predictions done by this model are correct. When the model generated by C4.5 predicts NI, it covers 92.90% of those cases but only 69.27% are correct predictions for the class of NI. J48 which is an improved version of C4.5 has better accuracy compared to the rest of the algorithm with 75.3% and it is the best result obtained in the family of tree. The outcome is most likely NI and the model confidence for this decision is only 66.67%. When the model predicts NI, it covers 82.78% of those cases. And it is correct with 78.35% of all predictions for class NI.

However, we found different case with PART. The predictive model generated by PART is super-confident that the correct prediction is for the class of Meet Expectation (ME). The confidence for this decision is high with 96.15%. The biggest support for this decision is coming from position or rank in the academic. Hemmings& Kay (2010) states that one of the important factor associated with qualifications and frequently linked to research output is academic rank. Blackburn & Lawrence (1995) and Smeby& Try (2005) also have generally reported a positive correlation between academic rank and publication output. Please keep in mind that 74.66% of all predictions done by this PART model are correct. When the model predicts ME, it covers 72.39% of those cases. And it is correct with 65.26% of all predictions for class ME.

Based on the PART model, the research found out that several factors had a great effect on publication output. Below are of the most effective factors for predicting ME:

1. Academic Rank (e.g. Professorship).
2. Number of Research Grant as Member (Active + Completed Project),
3. Number of Masters Student (Supervised)
4. Number of Research Grant as the Head of Project (Active + Completed Project).
5. Number of Dependants.
6. Number of Conference Attendance as Presenter.
7. Number of Leaves at the Beginning of the Year (Entitlement per year + Accumulation).

It is an interesting pattern discoveries when comparing effective or support factors for predicting ME and NI.

Retrieval Number: B10260782S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

For predicting Need Improvement, the attribute of Number of Leaves Taken Yearly is one of the effective support factors but for predicting Meet Expectation, Number of Leaves at the Beginning of the Year (Entitlement per year + Accumulation) are considered. We believed that one of the reason is, those who are classified as ME will have positive correlation or larger Number of Leaves at the Beginning of The Year compared to NI.

V. CONCLUSION

This research focus on the development of predictive models for predicting and evaluating academic staff performance in the function of performance management. The second objective of this study was to investigate the factors associated with scholarly publication productivity in UKM by using data mining to find unseen influence of attributes (factors) associated with scholarly publication productivity among academic staff. In the obtained dataset from the UKM Centre for Information Technology, this study mined the dataset hoping to find unseen patterns or relationship between personal factors such as age, administrative position, gender, number of students, number of research funds & annual performance appraisal score and publication research output among academic staff.

Based on the analysis of the results, the study concluded that the researcher's experience as a speaker at conference, number of research grant both as member and head of project was the variable that had the most influence on research output besides the researcher's academic position. Since the study was conducted only with records of Professor and Associate Professor, the research has not conclude yet and will move to the next phase of this study to generalize it for the larger sample which consist of all academic staffs from this institution. Further research can therefore be performed for all academic staffs of the university or at other public universities in the country to consider the differences in the working environments of those institutes. Finally, investigations into the other factors associated with a scholarly publication productivity need to be carried out to further determine the output efficiency. Analysis on predictive model of academic staff performance is an important process for both human resource and the Deanery of each faculty in a university, since it can help a better understanding of what will happen and support decision maker to intervene with corrective measures and ultimately improve it.

ACKNOWLEDGMENT

This research was supported by UniversitiKebangsaan Malaysia under the KRA 4 Research Grant (KRA-2018-014 & KRA-2018-015). We thank our colleagues from the Human Resource Departments and UKM Centre for Information Technology (PTM) who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

REFERENCES

[1] Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. Psychological Review, 84(2), 191.

[2] Beals, V. (2018). Predictive Analytics. In Webopedia. Retrieved from https://www.webopedia.com/TERM/P/predictive_analytics.html

[3] Blackburn, R. T., & Lawrence, J. H. (1995). Faculty at work: Motivation, expectation, satisfaction. Johns Hopkins University Press.

[4] Bland, C. J., Center, B. A., Finstad, D. A., Risbey, K. R., & Staples, J. G. (2005). A theoretical, practical, predictive model of faculty and department research productivity. Academic Medicine, 80(3), 225–237.

[5] Bland, C. J., Seaquist, E., Pacala, J. T., Center, B., &Finstad, D. (2002). One school’s strategy to assess and improve the vitality of its faculty. Academic Medicine, 77(5), 368–376.

[6] Brown, M. S. (2015). What IT Needs To Know About The Data Mining Process.

[7] Chien, C.-F., & Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with Applications, 34(1), 280–290. https://doi.org/10.1016/J.ESWA.2006.09.003

[8] Decision Tree. (n.d.). Retrieved November 20, 2018, from https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html

[9] Dhillon, S. K., Ibrahim, R., &Selamat, A. (2015). Factors associated with scholarly publication productivity among academic staff: case of a Malaysian public university. Technology in Society, 42, 160–166.

[10] Dunphy, D. C., & Hackman, B. K. (1988). Performance Appraisal as a Strategic Intervention. Asia Pacific Journal of Human Resources, 26(2), 23–34. https://doi.org/10.1177/103841118802600202

[11] Eckerson, W. W. (2007). Predictive analytics. Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report, 1, 1–36.

[12] Epstein, N., & Fischer, M. R. (2017). Academic career intentions in the life sciences: Can research self-efficacy beliefs explain low numbers of aspiring physician and female scientists? PloS One, 12(9), e0184543.

[13] Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization.

[14] Hemmings, B., & Kay, R. (2010). University lecturer publication output: Qualifications, time and confidence count. Journal of Higher Education Policy and Management, 32(2), 185–197.

[15] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2009). Knowledge discovery techniques for talent forecasting in human resource application. World Academy of Science, Engineering and Technology, 50, 775–783.

[16] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm. International Journal on Computer Science and Engineering, 2(8), 2526–2534.

[17] Kirimi, J. M., &Moturi, C. A. (2016). Application of Data Mining Classification in Employee Performance Prediction. International Journal of Computer Applications (0975-8887), 146(7).

[18] MacLennan, T. (2005). Moneyball: The Art of Winning an Unfair Game. The Journal of Popular Culture, 38(4), 780–781.

[19] Maltby, F., Gage, N. L., Berliner, D. C., & David, C. (1995). Educational psychology: an Australian and New Zealand perspective. John Wiley.

[20] Mueller, C. E., Gaus, H., &Konradt, I. (2016). Predicting Research Productivity in International Evaluation Journals across Countries. Journal of MultiDisciplinary Evaluation, 12(27), 79–92.

[21] Piatetsky-Shapiro, G., Fayyad, U., & Smith, P. (1996). From data mining to knowledge discovery: An overview. Advances in Knowledge Discovery and Data Mining, 1, 35.

[22] Reyes-Cruz, M. del R., &Perales-Escudero, M. D. (2016). Research self-efficacy sources and research motivation in a foreign language university faculty in Mexico: Implications for educational policy. Higher Education Research & Development, 35(4), 800–814.

[23] Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining,. Journal of Data Warehousing, 5(4), 13–22.

[24] Smeby, J., & Try, S. (2005). Departmental contexts and faculty research activity in Norway. Research in Higher Education, 46(6), 593–619.

[25] Turk, K., &Killumets, E. (2014). Performance Management of Academic Staff on the Example of the Faculties of Economics in University of Tartu and in Tallinn University of Technology.

[26] Waller, M. A., & Fawcett, S. E. (2013). Click here for a data scientist: Big data, predictive analytics, and theory development in the era of a maker movement supply chain. Journal of Business Logistics, 34(4), 249–252.