DATA MINING AND PREDICTIVE ANALYSIS ON THE EMPLOYMENT OF FRESH GRADUATE STUDENTS IN PUBLIC UNIVERSITIES IN MALAYSIA

NOR AZZIATY BINTI ABDUL RAHMAN

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

SULTAN IDRIS EDUCATION UNIVERSITY

2020

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

DATA MINING AND PREDICTIVE ANALYSIS ON THE EMPLOYMENT OF FRESH GRADUATE STUDENTS IN PUBLIC UNIVERSITIES IN MALAYSIA

NOR AZZIATY BINTI ABDUL RAHMAN

DISSERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE MASTER OF SCIENCE (INFORMATION SYSTEM AND MANAGEMENT)

FACULTY OF ART, COMPUTING & CREATIVE INDUSTRY
SULTAN IDRIS EDUCATION UNIVERSITY

2020

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

Please tick (√)
Project Paper
Masters by Research
Master by Mixed Mode
PhD

√

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on the 10 day of NOV 20 20

i. Student's Declaration:

I, NOR AZZIATY BINTI ABDUL RAHMAN (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled DATA MINING AND PREDICTIVE ANALYSIS ON THE EMPLOYMENT OF FRESH GRADUATE STUDENTS IN PUBLIC UNIVERSITIES IN MALAYSIA is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

Signature of the student

ii. Supervisor's Declaration:

I Dr. Mohd Hishamuddin bin Abdul Rahman (SUPERVISOR'S NAME) hereby certifies that the work entitled DATA MINING AND PREDICTIVE ANALYSIS ON THE EMPLOYMENT OF FRESH GRADUATE STUDENTS IN PUBLIC UNIVERSITIES IN MALAYSIA (TITLE) was prepared by the above named student, and was submitted to the Institute of Graduate Studies as a * partial/full fulfillment for the conferment of MASTER OF SCIENCE INFORMATION SYSTEM AND MANAGEMENT (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

8/12/2020
Date

DR. MOHD HISHAMUDDIN BIN ABDUL RAHMAN
Ketua Jabatan Komputeran
Fakulti Seni, Komputeran dan Industri kreatif
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak

Signature of the Supervisor

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: DATA MINING AND PREDICTIVE ANALYSIS ON THE EMPLOYMENT OF FRESH GRADUATE STUDENTS IN PUBLIC UNIVERSITIES IN MALAYSIA

No. Matrik /*Matric's No.*: M20161000828

Saya / *I* : NOR AZZIATY BINTI ABDUL RAHMAN

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.
4. Sila tandakan (√) bagi pilihan kategori di bawah */ Please tick (√) for category below:-*

☐	**SULIT/*CONFIDENTIAL***	Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. */ Contains confidential information under the Official Secret Act 1972*
☐	**TERHAD/*RESTRICTED***	Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. */ Contains restircted information as specified by the organization where research was done.*
☑	**TIDAK TERHAD / *OPEN ACCESS***	

(Tandatangan Pelajar/ Signature)

DR. MOHD HISHAMUDDIN BIN ABDUL RAHMAN
Ketua Jabatan Komputeran
Fakulti Seni, Komputeran dan Industri kreatif
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak

(Tandatangan Penyelia / *Signature of Supervisor*)
& (Nama & Cop Rasmi / *Name & Official Stamp*)

Tarikh: 8 DISEMBER 2020

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGMENT

"In the name of Allah, the Most Gracious and the Most Merciful"

Foremost, I would like to express my sincere gratitude to Dr. Tan Kian Lam and Dr. Lim Chen Kim, Faculty of Art, Computing and Creative Industry, Sultan Idris Education University, Tanjong Malim, Perak Malaysia as the supervisor and co-supervisor of this thesis for guidance and patience in giving me guidance during the implementation and preparation of this undergraduate thesis.

A special thanks to all the lecturers who have provided me with guidance and knowledge during my studies, the staffs of the Faculty of Art, Computing and Creative Industry, Sultan Idris Education University, the staffs of the Institute of Graduate Studies, Sultan Idris Education University who have helped directly or indirectly and who have shared their ideas.

I acknowledge the cooperation given by the Tracer Study Unit, Ministry of Higher Education, Malaysia for the cooperation given in providing graduate data for the purpose of this study.

Finally, thanks to my husband, children, mother and the whole family for their continued encouragement and support and always understand of my circumstances during preparing this thesis.

Thank you.

ABSTRACT

This study was aimed to use data mining to predict the employment of fresh graduate students in public universities in Malaysia. The research design of the study was model development using Rapid Miner Studio. The model used both supervised and unsupervised machine learning algorithms including k-Nearest Neighbor (kNN), Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine (SVM) and Neural Network. The sample consisted of 16,729 fresh graduate students were collected from the Tracer Study Unit of Ministry of Higher Education (MOHE). In order to build the classification model, Cross Industry Standard Process for Data Mining **(**CRISP-DM) methodology was applied. For the evaluation, 70% of the dataset were used as the training set and the remaining 30% were used as a testing set. To determine the error rate and to justify the accuracy of the proposed model objectively, classification error was used as the evaluation metric. The key finding of the predictive analysis revealed that employability among fresh graduate students can be predicted with 59.90% accuracy with a Neural Network as the most accurate predictive model. The significant factors contributing to graduates' employment were problem-solving and decision making skills. The unemployment, on the other hand, was mainly attributed to these factors – poor English competency, majoring in Malay, Education, and Science fields. In conclusion, the empirical data supported Neural Network model for predicting the employability among fresh graduate students in which the graduates should possess critical skills such as problem-solving and decision making skills. In implication, the predictive model was useful for graduate students, management of public institutions, Ministry of Higher Education, human resource personnel and academic staff in predicting the graduates' employability.

PERLOMBONGAN DATA DAN ANALISIS RAMALAN TERHADAP PEKERJAAN PELAJAR SISWAZAH DI UNIVERSITI-UNIVERSITI AWAM DI MALAYSIA

ABSTRAK

Kajian ini bertujuan menggunakan data mining untuk meramalkan pengambilan pelajar siswazah baharu di universiti-universiti awam di Malaysia. Reka bentuk kajian adalah pembangunan model menggunakan Rapid Miner Studio. Model ini menggunakan algoritma pembelajaran mesin yang diselia dan tidak diselia termasuk k-Nearest Neighbor (kNN), Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine (SVM) dan Neural Network. Sampel terdiri daripada 16,729 pelajar siswazah baharu yang dikumpulkan dari Unit Kajian Pengesanan Graduan, Kementerian Pengajian Tinggi (KPT). Untuk membina model klasifikasi, metodologi Cross Industry Standard Process for Data Mining (CRISP-DM) diterapkan. Untuk penilaian, 70% set data digunakan sebagai set latihan dan 30% selebihnya digunakan sebagai set ujian. Untuk menentukan kadar kesalahan dan untuk membenarkan ketepatan model yang dicadangkan secara objektif, kesalahan klasifikasi digunakan sebagai metrik penilaian. Penemuan utama analisis ramalan menunjukkan bahawa kebolehpasaran di kalangan pelajar siswazah baharu dapat diramalkan dengan ketepatan 59.90% dengan Neural Network sebagai model ramalan yang paling tepat. Faktor penting yang menyumbang kepada pekerjaan graduan adalah kemahiran menyelesaikan masalah dan membuat keputusan. Pengangguran, sebaliknya, disebabkan terutamanya oleh faktor-faktor ini - kecekapan Bahasa Inggeris yang lemah, jurusan Bahasa Melayu, Pendidikan, dan Sains. Kesimpulannya, data empirikal menyokong model Neural Network untuk meramalkan kebolehpasaran di kalangan pelajar siswazah baharu di mana siswazah harus mempunyai kemahiran kritikal seperti kemahiran menyelesaikan masalah dan membuat keputusan. Secara implikasinya, model ramalan berguna untuk pelajar siswazah, pengurusan institusi awam, Kementerian Pengajian Tinggi, kakitangan sumber manusia dan staf akademik dalam meramalkan kebolehpasaran graduan.

TABLE OF CONTENTS

CHAPTER 2 LITERATURE REVIEW

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network model
AODEsr	Averaged One-Dependence Estimators with subsumption resolution Cross Industry Standard Process for Data Mining (CRISP-DM)
AUDIT	Alcohol Use Disorders Identification Test
BAS	Behavioral Activation System
CDSE	Career decision self-efficacy
CFA	Confirmatory Factor Analysis
CGPA	Cumulative grade point average
DAX	Driving anger expression inventory
DV	dependent variable
EFA	Exploratory Factor Analysis
HEI	Higher Education Institutions
HEXACO	Honesty–Humility, Emotionalit, Extraversion, Agreeableness, Conscientiousness, and Openness
IPIP-50	International Personality Item Pool-50
LMS	Learning management system
MOHE	Ministry of Higher Education
Mturk	Amazon's Mechanical Turk2
PLS	Partial least squares (PLS)

POTKI	project proprietary study of implied knowledge inventory officers
PUKF	Pearson based on kernel functions
SEM	Structural Equation Modelling
SMO	Sequential Minimal Optimization
SMPY	Study of the Mathematically Precocious Youth
SN	Subjective Norm
STEM	Science, Technology, Engineering and Mathematics
SVM	Support Vector Machine
TAM	Technology Acceptance Model
TPB	Theory of Planned Behavior
UKM	Universiti Kebangsaan Malaysia
UM	University of Malaya
UPM	Universiti Putra Malaysia
USM	University of Science Malaysia
UTM	Universiti Teknologi Malaysia

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

This chapter discusses the introduction of this research. Section 1.1 presents the introduction, while Section 1.2 presents background of the study, Section 1.3 presents research objectives, Section 1.4 presents research questions, Section 1.5 presents problem statements, Section 1.6 presents research benefits, Section 1.7 presents research limitations and Section 1.8 presents the thesis outlines.

1.1 Introduction

In managing big data to discover new useful knowledge, the process of data mining is usually used. The Data Mining process is widely used by the technology companies,

banking, supermarkets and crime agencies to predict something in the future or explain something that has happened. There are several techniques in data mining, such as classification, statistical analysis, regression, prediction, evaluation, clustering and trend analysis, association analysis and etc.

There are two types of analysis commonly used, such as predictive analysis and descriptive analysis. Examples of predictive analyses are classification and regression. Classification and regression differ based on the outcomes either in a form of categorical or numerical. Descriptive analysis is different from predictive analysis. Descriptive analysis uses data mining and aggregation data to summarize what has happened or in other words is an analysis describing the past while predictive analysis uses modeling, data mining, machine learning techniques and various statistics to dig the historical data to determine what is possible occurring in the future because predictive analysis is probabilistic. Classification and prediction are the most important tasks in data mining. The selection of the algorithm often depends on the type of data (i.e. nominal, ordinal, ratio or interval) used. Machine learning was provided for every data mining algorithms and different data mining algorithms were used to set data based on knowledge.

This study proposed a suitable predictive model based on data mining techniques by using samples from Tracer Study, Ministry of Higher Education. The predictive model was proposed to generate attributes from students' records that affect the employment status of fresh graduates in public universities in Malaysia. This study consists of six chapters which discussed all the processes starting from

identifying the reseach problems until the discussion of the research questions or research objectives.

1.2 Background of the study

The issue of unemployment, especially among graduates, is an issue raised every year with approximately 200,000 graduates. One-fourth of graduates mostly with degree qualifications are unemployed after half a year of graduation (Michelle, 2016). It is very important to study the readiness among graduates to embark in the working world in order for local universities to produce graduates who are fully qualified with the needs of job industries by planning, researching and developing appropriate assessment tools.

Tracer Study Unit, Ministry of Higher Education, Malaysia had conducted a survey among fresh graduates to investigate trends of job evaluations: further study, upgrading skills, waiting for job placement, employed or unemployed. According to an online survey conducted by the Ministry of Higher Education in 2015 involving 107,850 respondents from a bachelor's degree, it was discovered that 58.0% of first degree holders were employed, 5.6% had further their studies, 2.2% were upgrading their skills, 6.2% were waiting for work placement and 27.9% were still unemployed. From the number of those who were employed, 11.3% were part-time workers while 88.7% were full-time workers (Kementerian Pengajian Tinggi Malaysia, 2015).

According to the statistics for the year 2015 released by National Education, Ministry of Higher Education, 27,065 out of 54,852 graduates in Higher Education Institutions (HEIs) were unemployed. It shows that graduates from the field of science contributed the highest rate of unemployment while graduates from the field of arts and social sciences contributed the lowest rate of unemployment in which the employment rate increased from 26.7% in 2015 to 25.6% in 2014.

There are six main areas of studies which contribute to an unemployment: applied sciences, human resource management, business administration, social sciences, accounting and arts, as reported by the Minister of Higher Education, Datuk Seri Idris Jusoh based on the statistics released by the Graduates Survey Detection System. According to the statistics, a total number of 54,103 unemployed graduates as compared to 238,187 graduates who graduated in 2016. However, employability among graduates increased slightly about 1.2 percent in 2016 compared to 2015. Various initiatives have been undertaken by the ministry to produce competitive and skillfull graduates. One of the initiatives is to introduce a work-based learning and introduce cumulative grade point average (iCGPA) at university level. The initiative introduced by the ministry is based on the Malaysian Education Blueprint 2015-2025 (Higher Education) which encourages higher institutions to produce holistic graduates ready to embark into fields of employment.

According to the statement from the SEEK Asia Chief Officer, Jake Andrew, graduates should have transferable skills and not just specialized skills to fulfill the demands of multipurpose jobs. Transferable skills such as interpersonal skills,

communications skills, organizational skills and leadership skills are the skills acquired for the working environment. Helpful, conflicts solving and internal motivation are interpersonal skills while public speaking, advice and persuade or sell are examples of communications skills. Additionally, setting and achieving goals, time management and multitasking are some examples of organizational skills while some examples of leadership skills include making decision, motivating others and working in groups.

Most organizations find that employees who have good qualifications, great personality won the favor of employers. According to a survey conducted by JobStreet.com in October 2016 involving 568 respondents from Human Resource Professionals in various industries, 14% of the respondents preferred qualifications while the majority of 51% prefer personality as the main criteria for selecting employees. There are five top reasons why fresh graduates are not employed as shown in Figure 1.1. According to a survey conducted by the Jobstreet.com in Nov 2015, the findings revealed the following results on reasons why fresh graduates were not employed: 59% stated the reasons were poor attitudes, characters or personalities, 60% said poor in communication skills, 60% said of being choosy of jobs/companies, 64% said poor command in English and 68% said demand for for unrealistic salary/benefits. This survey involved a total of 472 clients, managers and senior managers of Jobstreet.com in various industries in Malaysia. This study linked the attributes of the dataset with all five skills required to fulfill the needs of employers by using Data Mining and Predictive Analysis techniques. The results from the

analysis could be used by various organizations in planning the graduate job opportunities after graduation as summarized in subsection 1.6.

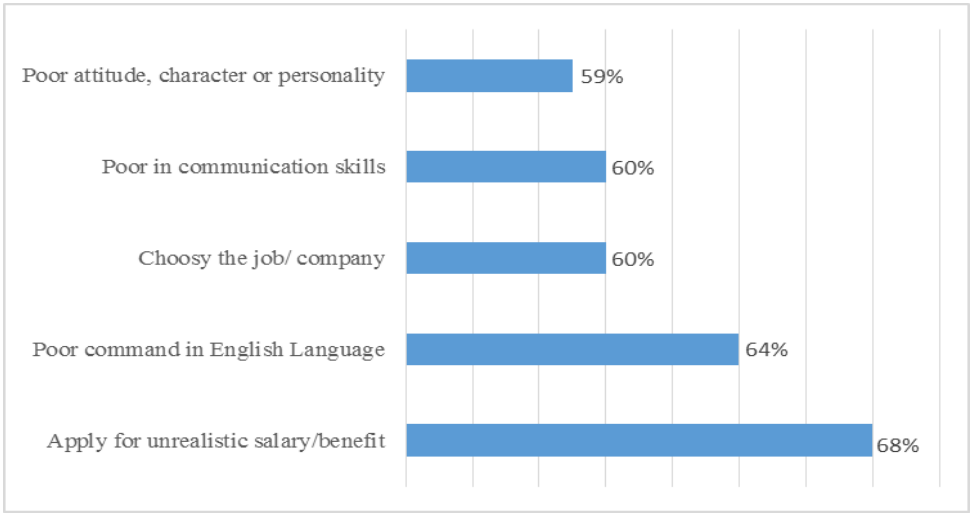

Figure 1.1. Five Top Reason for Fresh Graduate Students Do Not Get Hired (Jobstreet.com survey in November 2015)

1.3 Problem Statements

The main goal of all HEIs is to produce quality graduates from various aspects such as academic, personality, skills and able to meet the requirements of job industries. However, HEIs in Malaysia are facing problems and challenges in producing graduates who are skillfull because of the lack of assessments, information, prediction tools and analysis. HEIs has used the best pedagogy but still facing problems of unemployment among graduates. The process of analysing and understanding factors that can improve academic performances and skills was a complex process and need to correlate information in the past with the present to predict future graduates. Without good planning and management, the goal will not be achieved. This will lead

to dumping of graduates in the market which leads to the increasing of unemployment rate every year.

There are a large number of student data stored by the management of HEIs including the personal information, academic, co-curricular as well as the employment information but were not used in decisions and policies to improve the quality and performance of the students. The limited job opportunities are feared to be unable to accommodate the increasing number of graduates every year. Placement of graduates in industries is challenging due to a number of factors, such as academic performance, communication skills, capabilities, problem solving skills and disabilities. The studies of employability among Malaysian graduates are very limited although statistics from the Ministry of Higher Education (Kementerian Pendidikan Tinggi, 2016) showed a large number of graduates in local institutions who were unemployed after six months of graduation. Previous studies only examine the predictions of employability related to perceptions of career distress, factors affecting graduates engaging in their own business and employment performance.

It is evident from the discussion above that there is a need for studies to be conducted to clarify the gaps in relevant literature: type of samples used and limited studies in the area of employability, especially in Malaysia. The last gap is related to the lack of techniques used in analyzing skills and being competitive among the graduates before embarking into the job environment.

As a conclusion, there are three problems act as the basis of review of previous research:

1. There is a large amount of students' data but not used in making policies and decisions to improve the quality and performances of students especially in the placement of students in industries.
2. There are many existing attributes of the employability but no relationships are derived from these attributes.
3. Lack of predictive model in producing graduates who can meet the criteria of job selection by employers.

1.4 Research Questions

1. What are the factors that significantly cause fresh graduate students in Public Universities to be employed or unemployed?
2. What are the relationship within the attributes of employability among fresh graduate students in Public Universities?
3. What is the suitable classification model that can be used to predict trend of fresh graduate students in Public Universities to be employed or unemployed?

1.5 Research Objectives

The objectives of the study are as follows:

1. To investigate the existing factors that significantly causes fresh graduate students in Public Universities to be employed or unemployed.
2. To identify the relationship within the attributes of employability among fresh graduate students in Public Universities.
3. To develop the suitable classification model in predicting trend whether the fresh graduate students in Public Universities will be employed or unemployed.

1.6 Research Benefits

This study used data mining and predictive analysis to detect the marketability of graduate students from public universities in Malaysia. The benefits of the study are as follows:

1. The findings could help fresh graduate students to predict their future either employed or unemployed.
2. From the findings, graduates can also determine the attributes that can affect the graduates employed or unemployed.
3. The findings also may help the management of Public Institutions in assessing the success of graduates to be employed.

4. The findings also may help the Ministry of Education to assess the effectiveness of the courses offered by the university.
5. The findings may also help human resources to make projections of labor each year.

1.7 Research Limitations

The study involved samples data from five Public Universities in Malaysia for the years 2015, Universiti Teknologi Malaysia (UTM), Universiti Sains Malaysia (USM), Universiti Putra Malaysia (UPM), University of Malaya (UM) and Universiti Kebangsaan Malaysia (UKM). The data was developed by the Tracer Study, Ministry of Higher Education, Malaysia. The data used in this study only involved five public universities because this is the only data provided and authorized by the Tracer Study Unit. Therefore, the findings depends on the sincerity of the respondents to provide feedback on a questionnaire in Tracer Study during the graduation.

1.8 Thesis Outline

The remainder of this thesis is structured as Chapter 2 presents the literature review about data mining and predictive analysis, Chapter 3 presents the overall methodology, Chapter 4 proposed the solutions, Chapter 5 presents the evaluations, results and discussion and Chapter 6 presents the contributions and future works.